TR-NLP-UT-2006-1*

GENIA Corpus Manual

Encoding schemes for the corpus and annotation

Jin-Dong Kim[1,3,4], Tomoko Ohta[1], Yuka Teteisi[2], Jun'ichi Tsujii[1,3,4]

[1]University of Tokyo,
[2]Kogakuin University,
[3]University of Manchester
[4]National Centre for Text Mining

Date of Release: 15 November 2006

---

* All technical reports from Tsujii Laboratory can be found at http://www-tsujii.is.s.u-tokyo.ac.jp/TR/.

1. GENIA corpus

The GENIA corpus a product of the GENIA project of which the objective is to develop information extraction (IE) and text mining (TM) systems for the specific subject domain of molecular biology and medical science. The purpose of developing the corpus is to provide reference materials for training and evaluation of the IE and TM systems.

1.1. Compilation

Currently, the GENIA corpus is made up of the titles and abstracts of journal articles which have been taken from the Medline database. Whereas the Medline indexes broad range of academic articles covering the general or specific domains of life sciences, GENIA is intended to cover a smaller subject domain: biological reactions concerning transcription factors in human blood cells. To retrieve the corresponding records from Medline, we used the following search query on the PubMed web interface: “Humans”[MeSH] AND “Blood Cells”[MeSH] AND “Transcription Factors”[MeSH].

1.2. Annotation

The GENIA corpus has been being annotated by encoding human’s interpretation of the text. The purpose of the annotation is two fold. First, we annotate the corpus to make the biomedical knowledge encoded in the text transparent. Second, we also annotate the corpus to reveal the syntactic structure behind the text. Eventually, our objective is to establish the mapping between the knowledge pieces and the linguistic structures.

2. Document Encoding

Currently, the source of the GENIA corpus is the Medline database. Each Medline record can be retrieved from PubMed by using its PMID (PubMed ID). Accordingly, a Medline record may have its URI like as follows: http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=7641692.

The above URI represents the Medline record whose PMID is 7641692. It is an actual URL by which the Medline record encoded in XML can be retrieved. Figure 2-1 shows the retrieved XML record. Note that it only shows selected parts of the documents with the other parts collapsed into ellipsis symbols.

```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st November 2004//EN"
"http://www.ncbi.nlm.nih.gov/entrez/query/DTD/pubmed_041101.dtd">
<PubmedArticleSet>
<PubmedArticle>
  <MedlineCitation Owner="NLM" Status="MEDLINE">
    <PMID>7641692</PMID>
    …
    <Article PubModel="Print">
      …
      <ArticleTitle>IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.</ArticleTitle>
      …
      <Abstract>
        <AbstractText>Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation. In primary T lymphocytes we show that CD28 ligation leads to the rapid intracellular formation of reactive oxygen intermediates (ROIs) which are required for CD28-mediated activation of the NF-kappa B/CD28-responsive complex and IL-2 expression. Delineation of the CD28 signaling cascade was found to involve protein tyrosine kinase activity, followed by the activation of phospholipase A2 and 5-lipoxygenase. Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation. These findings should be useful for therapeutic strategies and the development of immunosuppressants targeting the CD28 costimulatory pathway.</AbstractText>
      </Abstract>
      …
    </Article>
    …
  </MedlineCitation>
  …
</PubmedArticle>
</PubmedArticleSet>
```

Figure 2-1. A Medline record encoded in XML

Since a Medline record is huge including a lot of meta data, we do not take the entire record. Instead, we take only the elements containing natural language texts and include them in the corpus according to the "GENIA Corpus Curation Framework". Figure 2-2 shows a file of GENIA corpus. The corpus file has two **excerpt** elements containing text from the **ArticleTitle** and **AbstractText** elements of the original XML file. Each excerpt element maintains all the information required to identify the source of the text.

<?xml version="1.0"?>
<excerptset>
<excerpt id="25" srcDB="Medline" srcrecID="7641692" date="08SEP2006" texttype="article-title" seqnext="26" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=7642692#xpointer(/PubmedArticleSet[1]/PubmedArticle[1]/MedlineCitation[1]/Article[1]/ArticleTitle[1])">
IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.**</excerpt>**
<excerpt id="26" srcDB="Medline" srcrecID="7641692" date="08SEP2006" texttype="abstract" seqprev="25" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=7641692#xpointer(/PubmedArticleSet[1]/PubmedArticle[1]/MedlineCitation[1]/Article[1]/Abstract[1]/AbstractText[1])">Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation. In primary T lymphocytes we show that CD28 ligation leads to the rapid intracellular formation of reactive oxygen intermediates (ROIs) which are required for CD28-mediated activation of the NF-kappa B/CD28-responsive complex and IL-2 expression. Delineation of the CD28 signaling cascade was found to involve protein tyrosine kinase activity, followed by the activation of phospholipase A2 and 5-lipoxygenase. Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation. These findings should be useful for therapeutic strategies and the development of immunosuppressants targeting the CD28 costimulatory pathway.**</excerpt>**
</excerptset>

Figure 2-2. A GENIA corpus file

3. Sentence Segmentation

Sentence segmentation is regarded as one of the initial steps of natural language processing, segmenting an input text into sentences. Having text segmented into sentences, further analysis is usually performed sentence by sentence. For sentence segmentation, GENIA uses an XML tag, **sentence** which has unique identifier **id** for maintenance purposes. Figure 3-1 shows an excerpt with text segmented into sentences. Note that annotation of GENIA including the sentence segmentation involves only insertion of XML mark-ups and does not involve any other modification of original text including white spaces, so that the original text can be restored at any time by simply dropping the mark-ups.

<excerpt id="26" srcDB="Medline" srcrecID="7641692" date="08SEP2006" texttype="abstract" seqprev="25" srctxtURL="http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&rettype=medline&id=7641692#xpointer(/PubmedArticleSet[1]/PubmedArticle[1]/MedlineCitation[1]/Article[1]/Abstract[1]/AbstractText[1])"><**sentence id="E26S1"**>Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation.</**sentence**> <**sentence id="E26S2"**>In primary T lymphocytes we show that CD28 ligation leads to the rapid intracellular formation of reactive oxygen intermediates (ROIs) which are required for CD28-mediated activation of the NF-kappa B/CD28-responsive complex and IL-2 expression.</**sentence**> <**sentence id="E26S3"**>Delineation of the CD28 signaling cascade was found to involve protein tyrosine kinase activity, followed by the activation of phospholipase A2 and 5-lipoxygenase.</**sentence**> <**sentence id="E26S4"**>Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation.</**sentence**> <**sentence id="E26S5"**>These findings should be useful for therapeutic strategies and the development of immunosuppressants targeting the CD28 costimulatory pathway.</**sentence**></excerpt>

Figure 3-1. A GENIA corpus file, sentence-segmented

4. Tokenization

Tokenization is one of the initial steps of natural language processing, preparing the most fundamental unit of processing. Tokenization is to change an input text from a string of characters to a sequence of tokens which are non-overlapped and non-recursive. Usually, a word, a group of characters delimited by whitespaces, becomes a token if it does not include a punctuation mark. If it does, the punctuation mark becomes a separate token. The treatment of punctuations may vary.

We generally follow the Penn Treebank tokenization scheme except some modification we have made. For the detail of the modification, readers are referred to the "GENIA tokenization & POS tagging guidelines" at http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/GENIA-POS-README.html.

For tokenization, GENIA uses an XML tag, **w** for both words and punctuation marks. Figure 4-1 shows a sentence which has been tokenized. Note that after the tokenization, all the spacing information has to be preserved so that the original text can be recovered simply by dropping the tokenization tags. It is a widely accepted convention to make token not include space characters.

<sentence><**tok**>IL-2</**tok**> <**tok**>gene</**tok**> <**tok**>expression</**tok**>
<**tok**>and</**tok**> <**tok**>NF-kappa</**tok**> <**tok**>B</**tok**> <**tok**>activation</**tok**>
<**tok**>through</**tok**> <**tok**>CD28</**tok**> <**tok**>requires</**tok**> <**tok**>reactive</**tok**>
<**tok**>oxygen</**tok**> <**tok**>production</**tok**> <**tok**>by</**tok**> <**tok**>5-lipoxygenase</**tok**><**tok**>.</**tok**></sentence>

Figure 4-1. A sentence with POS annotation.

5. Part-of-speech tagging

Part-of-speech tagging is another initial step of natural language processing which is often performed right after tokenization. After tokenization, every token is assigned a POS label. The GENIA POS annotation generally follows the Penn Treebank POS tagging scheme. For the modification we have made, readers are referred to the GENIA tokenization & POS tagging guidelines" at http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/GENIA-POS-README.html.

For POS labeling, GENIA uses an attribute **cat** (representing category) for the **tok** tag. When the POS is not uniquely determined, the **cats** (representing candidate categories) attribute is used to enumerate the candidate labels. Figure 5-1 shows a sentence with POS-labeled tokens. Note that the word *CD28* has two candidate POS labels enumerated as space-separated values in the attribute **cats**.

```
<sentence><tok cat="NN">IL-2</tok> <tok cat="NN">gene</tok> <tok
cat="NN">expression</tok> <tok cat="CC">and</tok> <tok cat="NN">NF-
kappa</tok> <tok cat="NN">B</tok> <tok cat="NN">activation</tok> <tok
cat="IN">through</tok> <tok cats="NN FRAG">CD28</tok> <tok
cat="VBZ">requires</tok> <tok cat="JJ">reactive</tok> <tok
cat="NN">oxygen</tok> <tok cat="NN">production</tok> <tok
cat="IN">by</tok> <tok cat="NN">5-lipoxygenase</tok><tok
cat="PERIOD">.</tok></sentence>
```

Figure 5-1. A sentence with POS annotation

Table 5-1 lists the POS labels used for GENIA POS annotation. For convenience, the corresponding PTB labels are also given when they are different from GENIA.

Table 5-1. Part-of-speech labels of GENIA and PTB

PTB / GENIA	Description
CC	Coordinating conjunction.
CD	Cardinal number.
DT	Determiner.
EX	Existential *there*.
FW	Foreign word.
IN	Preposition or subordinating conjunction.
JJ	Adjective.
JJR	Adjective, comparative.
JJS	Adjective, superlative.
LS	List item marker.
MD	Modal.
NN	Noun, singular or mass.
NNS	Noun, plural.
NNP	Proper noun, singular.
NNPS	Proper noun, plural.

PDT		Predeterminer.
POS		Possessive ending.
PRP		Personal pronoun.
PRP$	PRPP	Personal pronoun, possessive.
RB		Adverb.
RBR		Adverb, comparative.
RBS		Adverb, superlative.
RP		Particle.
SYM		Symbol.
TO		*to*.
UH	-	Interjection. This doesn't appear in the GENIA corpus.
VB		Verb, base form.
VBD		Verb, past tense.
VBG		Verb, present participle or gerund.
VBN		Verb, past participle.
VBP		Verb, non-3rd person singular present.
VBZ		Verb, 3rd person singular present.
WDT		*Wh*-determiner.
WP		*Wh*-pronoun.
WP$	WPP	*Wh*-pronoun, possessive.
WRB		*Wh*-adverb.
#	-	Pound sign. This doesn't appear in the GENIA corpus.
$	-	Dollar sign. This doesn't appear in the GENIA corpus.
.	PERIOD	Period.
,	COMMA	Comma.
:	COLON	Colon.
(	LRB	Left one of any paired symbols used as brackets: (, [,{, <.
)	RRB	Right one of any paired symbols used as brackets:),], }, >.
“	LQT	Left quotation mark, single or double
”	RQT	Right quotation mark, single or double

6. Syntactic Tree annotation

Generally, syntactic tree annotation reveals how words in a sentence are organized to form the meaning of the sentence. In a sentence, words may be grouped together into phrases. Again, phrases together with or without other words may also be grouped to form larger phrases or clauses. It may continue until eventually the whole sentence is grouped together. This analysis usually yields a tree of phrases and clauses with the root element covering the whole sentence, internal elements corresponding to phrases and clauses, and leaf elements corresponding to words.

The term constituent is used to refer to a word, a phrase or a clause which function as a unit and is embedded into a hierarchical structure. There are many theories and formalisms about how to group the words to form constituents, how to classify the constituents, and how to encode other useful information. The annotation scheme of the GENIA tree annotation (a.k.a GENIA Treebank, GTB) has been designed based on the Penn Treebank II (PTB) bracketing guidelines (Beis et al, 1995).

6.1. Chunking and Constituent Labeling

For syntactic annotation, GENIA uses an XML tag **cons** which have the **cat** attribute to specify the syntactic category of each constituent. Table 6-1 lists the syntactic categories for GTB.

Table 6-1. Syntactic categories of constituents

	Categories		Description
	PTB	GTB	
Clause level	S		Simple declarative sentence
	SBAR		Clause introduced by a subordinating conjunction
	SBARQ		Direct question introduced by a *wh*-word or *wh*-phrase.
	SINV		Sentence with subject-auxiliary inversion
	SQ		Question without *wh*-phrase
Phrase level	ADJP		Adjective phrase
	ADVP		Adverb phrase
	CONJP		Conjunction phrase
	FRAG		Fragment
	INTJ	-	Interjection
	LST		List item
	NAC	NP	Not a constituent
	NP		Noun phrase
	NX		Used to mark the head of the noun phrase
	PP		Prepositional phrase
	PRN		Parenthetical expression
	PRT		Particle
	QP		Quantifier phrase
	RRC		Reduced relative clause
	UCP		Unlike coordination phrase
	VP		Verb phrase

	WHADJP		*wh*-adjective phrase
	WHADVP		*wh*-adverb phrase
	WHNP		*wh*-noun phrase
	WHPP		*wh*-prepositional phrase
	X	-	Unknown, uncertain, or unbracketable
	-	COMP	Used to specify the null complementizer

They are almost the same as the constituent labels defined in PTB II bracketing guidelines, but a couple of modifications have been made.

- The PTBII Labels **NAC** (a prenominal modifier that is not a constituent) and **NX** (the head of a complex noun phrase) have been merged to NP (noun phrase).
- When a noun phrase consists of a sequence of nouns, the internal structure is left unstructured unless coordination is involved.

These two modifications are in order to simplify the process of annotation. In case of biomedical abstracts, long noun phrases often involve multi-word technical terms whose syntactic structure is difficult to determine without deep domain knowledge. On the other hand, the structure of the noun phrases are usually independent of the structure outside the phrase, so that it would be easier to analyze the phrases involving such terms independently (e.g. by biologists) and later merge the two analyses together. Thus we have decided that we leave noun phrases unstructured in GTB annotation unless their analysis is necessary for determining the structure outside the phrase.

6.2. Function Tags

The function tags indicate semantic roles and other related information not easily encapsulated in the simple constituent labels.

Grammatical roles

The grammatical roles described in the PTBII guidelines are encoded in GTB using the **role** and **top** attributes of the **cons** elements. The labels for grammatical roles are listed in the following table.

Attribute	Value	Corresponding construction
Role	SBJ	Surface subject
	LGS	Logical subject in passive
	DTV	Dative
	PRD	Predicate which is not VP
	VOC	Vocative
	PUT	Locative complement of *put* (currently unused)
Tpc	TPC	Topicalized element

The **TPC** label has been separated from the other 6 because there are chances for the TPC tag together with another grammatical tag to be assigned to the same structure.

Form/Function discrepancies

The form/function discrepancies described in the PTBII guidelines are encoded in GTB using the **func** attribute of the **cons** elements.

Attribute	Value	Corresponding construction
Func	NOM	Nominal
	ADV	Adverbial

Adverbials

Adverbials are generally VP adjunct describing action, event or situation from semantic aspects. The PTBII guidelines define 7 labels for adverbials, but for GTB currently only **TMP** is used.

Attribute	Value	Corresponding semantic type
Adv	BNF	Benefactive (unused)
	DIR	Direction (unused)
	EXT	Extent (unused)
	LOC	Locative (unused)
	MNR	Manner (unused)
	PRP	Purpose or reason (unused)
	TMP	Temporal expression

Miscellaneous

The PTBII guidelines define miscellaneous 4 labels among which **CLF** and **HLN** are currently used for GTB.

Attribute	Value	Corresponding construction
Misc	CLF	Cleft (*it is … that …*)
	HLN	Headline
	TTL	Title (unused)
	CLR	Closely related (unused)

For GTB the **syn** attribute is additionally prepared to specify the syntactic idiosyncrasy of constituents. Currently, it is only used to denote constituents involving coordination.

Attribute	Value	Corresponding construction
Syn	COOD	Coordination

Figure 6-1 shows examples of constituents involving coordination.

<sentence><cons cat=”S”><cons cat=”NP” role=” SBJ”><cons cat=”NP”><cons cat=”ADJP”>Structurally related </cons>divalent cations</cons> <cons cat=”PP”>like **<cons cat=”NP” syn="COOD">**<cons cat=”NP”>cobalt</cons>, <cons cat=”NP”>nickel</cons>, and <cons cat=”NP”>mercury</cons>**</cons>**</cons></cons> <cons cat=”ADVP”>also</cons> <cons cat=”VP”><cons cat=”ADVP”>partially</cons> increase <cons cat=”NP”>monokine secretion</cons> but <cons cat=”PP”>to <cons cat=”NP”>a **<cons cat=”ADJP” syn="COOD">**<ADJP>much lower</cons> and thus <cons cat=”ADJP”>insignificant </cons>**</cons>** extent</cons></cons></cons>.</cons></sentence>

Figure 6-1. Annotation for coordinated structures.

6.3. Null elements

A null constituent is marked as an empty element with the **null** attribute specifying the type of the null element. The syntactic category of null complementizers is denoted as COMP, yielding <cons cat=“COMP” null=”ZERO”/>.

Attr	Value		Usage
	PTB	GTB	
null	*T*	T	Trace of A’-movement, including parasitic gaps
	*	NONE	Arbitrary PRO, controlled PRO, and trace of A-movement
	0	ZERO	Null complementizer, including null *wh*-operator
	U	U	Unit (unused)
	?	QSTN	Placeholder for ellipsed material
	NOT	NOT	Anti-placeholder in template gapping (unused)
	RNR	RNR	Pseudo-attach: right node raising
	ICH	ICH	Pseudo-attach: interpret constituent here
	EXP	EXP	Pseudo-attach: expletive
	PPA	PPA	Pseudo-attach: permanent predictable ambiguity (unused)

6.4. Coindexing

When necessary, a constituent (a **cons** element) may be assigned a unique id as the value of the **id** attribute. The indexed constituents then may be referenced to coindex associated constituents. In PTBII and GTB, the coindexing mechanism is used to represent the association of null elements and gapped structure.

Null constituent coindexing

A null constituent (represented as an empty element) is coindexed with its associated constituent using the **ref** attribute. An example is in Figure 6-2.

<sentence><cons cat=“S”><cons cat=“PP”>In <cons cat=“NP”>the present paper</cons></cons>, <cons cat=“NP” role=”SBJ” **id="i55"**><cons cat=“NP”>the binding</cons> <cons cat=“PP”>of <cons cat=“NP”>a [125I]-labeled aldosterone derivative</cons></cons> <cons cat=“PP”>to <cons cat=“NP”><cons cat=“NP”>plasma membrane rich fractions</cons> <cons cat=“PP”>of HML</cons></cons></cons></cons> <cons cat=“VP”>was <cons cat=“VP”>studied <cons cat=“NP” null="NONE" **ref="i55"**/></cons></cons>.</cons></sentence>

(S (PP In (NP the present paper)), (NP-SBJ-**55** (NP the binding) (PP of (NP a [125I]-labeled aldosterone derivative)) (PP to (NP (NP plasma membrane rich fractions) (PP of HML)))) (VP was (VP studied *-**55**)).)

Figure 6-2. A tree-annotated sentence in XML and PTB formats

Gap coindexing

When a gapped construction exists alongside a complete (ungapped) clause of parallel structure within a sentence, the complete clause is used as a template to which elements in the gapped clause are mapped via the **map** attribute referencing the corresponding elements (referred to as “gap coindexing”). An example is in Figure 6-3.

<sentence><cons cat=“S”><cons cat=“PP” top="TPC" id="i75">Of <cons cat=“NP”>the 23 cases</cons></cons>, <cons cat=“S” syn="COOD"><cons cat=“S”>**<cons cat=“NP” role=”SBJ” id="i76">**<cons cat=“NP”>19</cons> <cons cat=“PP” null="T" ref="i75"/>**</cons>** <cons cat=“VP”>were <cons cat=“VP”>classified <cons cat=“NP” null="NONE" ref="i76"/> **<cons cat=“PP” id="i78">**as NK-cell**</cons>**</cons></cons></cons> and <cons cat=“S”>**<cons cat=“NP” role=”SBJ” map="i76">** <cons cat=“NP”>4</cons> <cons cat=“PP” null="T" ref="i75"/>**</cons>** **<cons cat=“PP” map="i78">**as <cons cat=“NP”>T-cell tumours</cons>**</cons>**</cons></cons>.</cons></sentence>

(S (PP-TPC-75 Of (NP the 23 cases)), (S-COOD (S (NP-SBJ-76 (NP 19) (PP *T*-75)) (VP were (VP classified (NP *-76) (PP-78 as NK-cell)))) and (S (NP-SBJ=76 (NP 4) (PP *T*-75)) (PP=78 as (NP T-cell tumours)))) .)

Figure 6-3. Gep coindexing

7. Term annotation

Term annotation is performed to identify expressions with a specific meaning which are shared and repeatedly used in a particular subject domain. For practical reasons, consideration is often limited to nominal expressions as they appear in more fixed form than verbal expressions. In GENIA term annotation, biological terms appearing in nominal expressions in literature are located and identified with concepts from ontologies while mostly covering names of biological entities. The identified terms are then expected to establish anchors for knowledge extraction.

Syntactically, a term is defined as a head noun plus some of preceding qualifiers as shown in Figure 7-1. The list of qualifiers to be included in a term is not fixed. Instead, the decision of inclusion or exclusion of a specific qualifier in or out of a term is left to the annotator's decision. It may seem arbitrary, but we are expecting we could collect statistics on experts' tendencies concerning the in/exclusion and based on the statistics we could set more elaborate scheme for the future annotations or revision.

<term> := <qualifier>* <head noun>
<qualifier> := <adjective>|<noun modifier>

Figure 7-1. The syntactic definition of terms

By the definition, a term is a group of consecutive tokens. Since we only consider nominal expressions and do not allow prepositional phrases inside a term, term will appear inside a noun phrase. In terms of granularity of annotation structures, the term is larger than the token and smaller than the constituent. In other words, a term will occur inside a cons (probably in a noun phrase) as a group of tokens.

Term annotation is performed in two stages: first, the text region corresponding to a term is marked as an XML element, **term**.; second, a semantic descriptor is assigned to the each term element as a value of **sem** attribute. The semantic descriptor is assumed to come from a the GENIA ontology which hierarchically classifies and defines biological concepts. Terms are allowed to be embedded in other terms.

Figure 7-2 shows an example of term annotation to the text "*IL-2 gene transcription in T cells*". The three terms, "*IL-2 gene*", "*IL-2 gene transcription*" and "*T cells*" are marked as term elements. To the terms, "*IL-2 gene*" and "*T cells*", the semantic descriptors, "*DNA_domain_or_region*" and "*Cell_type*" are assigned respectively. To the term IL-2 gene transcription", no descriptor is assigned since there are no corresponding definition in the GENIA ontology. Note that the term "*IL-2 gene*" is embedded in another term "*IL-2 gene transcription*".

<term><term sem="DNA_domain_or_region">**IL-2 gene**</term>
transcription</term> **in** <term sem="Cell_type">**T cells**</term>

Figure 7-2. GENIA term annotation – a simple case

In natural language text, sometimes, terms appear in discontinuous text regions. For example, in the text, "*CD2 and CD 25 receptors*", the term "*CD2 receptor(s)*" appear by two discontinuous text regions, "*CD2*" and "*receptors*", whereas another term "*CD25 receptor*" appears by a text region. Usually discontinuous terms appear when terms are coordinated and a part of the terms are shared by them. In GENIA, coordinated terms are annotated together as a **coterm** element when they have a shared part causing discontinuous terms. The sem attribute of the coterm element is then used to enumerate the semantic descriptors of involved terms plus the semantic descriptor of the coordinating connector. The participating term fragments are also marked up for further processing.

Figure 7-3 shows an example of term annotation to the sample text. The entire text is annotated as a **coterm** element referencing two protein molecules which are connected by the "AND" coordinator.

<coterm sem="(AND Protein_molecule Protein_molecule)"><frag>**CD2**</frag> **and** <frag>**CD25**</frag> <frag>**receptors**</frag></coterm>

Figure 7-3. GENIA term annotation for coordinated clauses

Figure 7-4 shows the GENIA term ontology. Among the concepts defined in the ontology, only the leaf concepts in the tree structure are used as semantic descriptor of the term annotation. For detailed description of the ontology, readers are referred to another technical report "GENIA ontology"

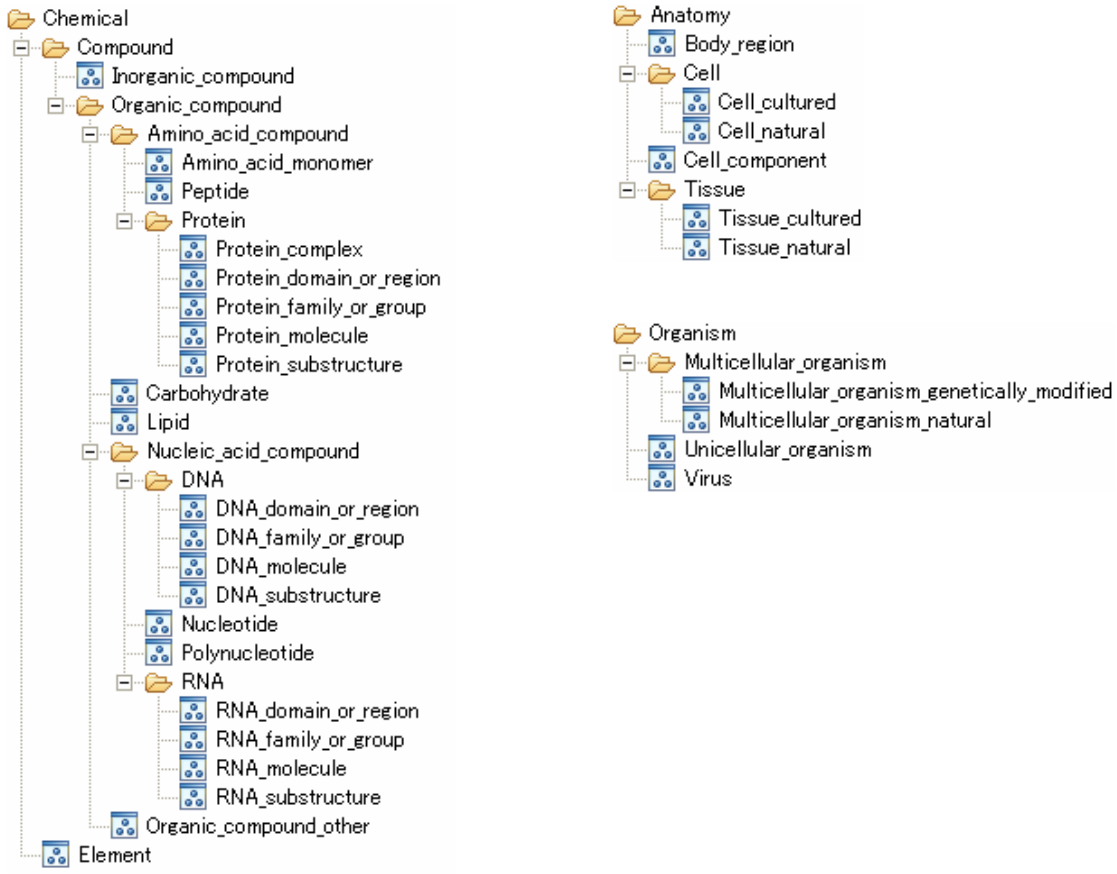


Figure 7-4. The hierarchy of the GENIA term ontology

8. Event annotation

To the GENIA corpus the event annotation is performed to identify biological events mentioned in natural language text. In GENIA, a biological *event* is defined as a temporal occurrence that happens to one or more biological entities. Especially, a number of events which cause some specific change on genes or gene products (proteins) are defined in the GENIA event ontology, becoming the target of GENIA event annotation. Figure 8-1 the hierarchy of the GENIA event ontology

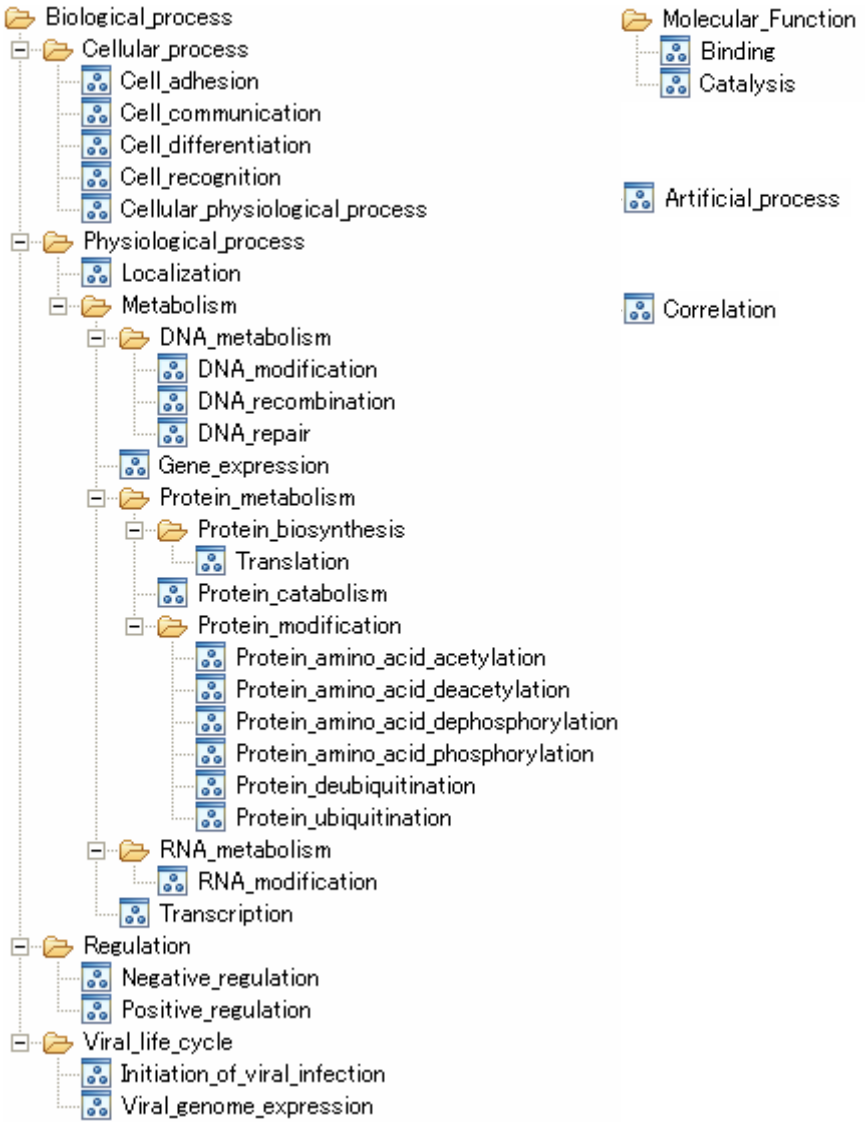


Figure 8-1. The Hierarchy of the GENIA event ontology

Table 8-1 shows some excerpts from Medline abstracts and events detected in them according to the GENIA ontology. The excerpts have been selected to demonstrate representative problems of event annotation. The PMID of the source article is given in the parenthesis next to each excerpt.

Table 8-1. Medline text excerpts and events mentioned in them

Ex. 1	Lipopolysaccharide induces phosphorylation of MAD3 … (PMID8505309)
	Event #1 Type : *Protein_amino_acid_phosphorylation* Theme : *MAD3* (Protein_molecule) Event #2 Type : Positive_regulation Theme : Event #1 Cause : *Lipopolysaccharide*
Ex. 2	Enhancement of human immunodeficiency virus 1 replication in monocytes by 1,25-dihydroxycholecalciferol. (PMID1650477)
	Event #3 Type : RNA_metabolism Theme : *human immunodeficiency virus 1* Event #4 Type : Positive_regulation Theme : Event #3 Cause : *1,25- dihydroxycholecalciferol*
Ex. 3	… I kappa B/MAD-3 binds directly to NF-kappa B p50 … (PMID1493333)
	Event #5 Type : Binding Theme : *I kappa B/MAD-3* and *NF-kappa B p50*
Ex. 4	Expression of M10 did not affect induction of HIV transcription … (PMID1402661)
	Event #6 Type : Gene_expression Theme : *M10* Event #7 Type : Transcription Theme : *HIV* Event #8 (negation) Type : Positive_regulation Theme : Event #7 Cause : Event #6
Ex. 5	In this report, we demonstrate that a novel Ets-related transcription factor, Elf-1, binds specifically to two purine-rich motifs in the HIV-2 enhancer. (PMID1527846)
	Event #9 Type : Binding

	Theme : *E1f-1* and *HIV-2 enhancer* Event #10 Type : Binding Theme : *E1f-1* and *two purine-rich motifs* (in HIV-2 enhancer)
Ex. 6	Similar to its effect on the induction of AP1 by okadaic acid, PMA inhibits the induction of c-jun mRNA by okadaic acid. (PMID1851743)
	Event #11 Type : Positive_regulation Theme : *c-jun mRNA* Cause : *okadaic acid* Event #12 Type : Negative_regulation Theme : Event #11 Cause : *PMA* Event #13 Type : Positive_regulation Theme : *AP1* Cause : *okadaic acid* Event #14 Type : Negative_regulation Theme : Event #13 Cause : *PMA*
Ex. 7	… HS-40 behaved as an authentic enhancer for high-level zeta 2 globin promoter activity … (PMID8455611)
	Event #15 Type : Positive_regulation Theme : *zeta 2 globin promoter* Cause : *HS-40*

In GENIA, an event is identified with its *themes* and *causes*. The objects undergoing change during an event are called themes and the objects causing the change are called causes of the event. By the excerpts in Table 8-1, the policy of GENIA event identification is described as follows:

- Most event classes relate to one theme:
  - Event #1, #2, #3, #4, #6, #7, #8, #11, #12, #13, #14, #15,
- but some relate to more than one themes:
  - Event #5, #9, #10.
- Usually, events act upon entities, making entities themes of the events:
  - Event #1, #3, #5, #6, #7, #9, #10, #11, #13, #15,
- but some events act upon other events:
  - Event #2, #4, #8, #12, #14.

- Some events are connected with causes:
    - Event #2, #4, #8, #9, #10, #11, #15.
- Usually, entities cause events:
    - Event #2, #4, #9, #10, #11, #15 ,
- but sometimes an event may be caused by another event:
    - Event #8.
- An event may be negated:
    - Event #8.

Figure 8-2 shows a sentence and event annotation made to the sentence in XML (a) and in a styled view using CSS (b). In GENIA XML encoding, a sentence may be followed by one or more event elements each of which encodes an event mentioned in the sentence. The event element encodes the type, themes and causes of an identified event. For the type of an event, a descriptor from the GENIA event ontology may be specified. For the themes and causes of an event, the IDs of pre-annotated terms or events may be referenced. Since the GENIA annotation seeks for linking of information pieces and language structures, the **clue** element has been prepared in the event element to reveal the text parts which participate in mentioning the event. Inside the **clue** element, the text spans which are responsible for mentioning the type of the event, linking the event type to the themes and linking the event type to the causes are marked-up as the **ClueType**, **LinkTheme** and **LinkCause** elements respectively.

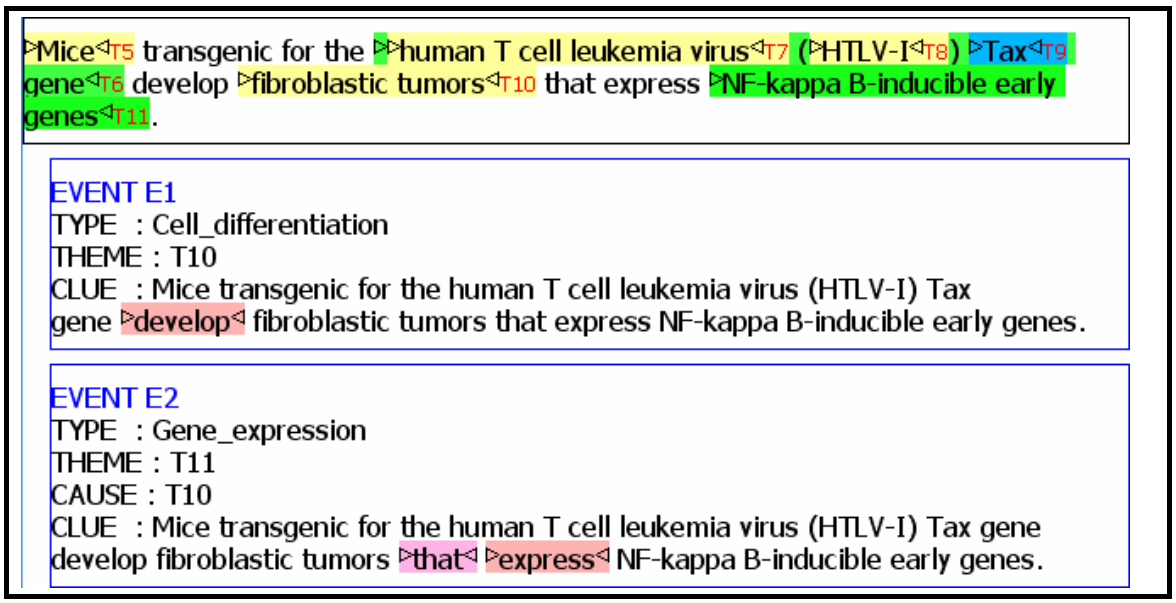
In the figure, the event E1 identifies the event of tumor development which has been classified as a Cell_differentiation event of which the theme is the *fibroblastic tumors*. The text span *develop* is determined to give a clue for the event classification.

The event E2 identifies an event of Gene_expression whose theme is *NF-kappa B-inducible early genes* and cause is *Mice*. The text span *express* has been determined to give a clue for determining the event class and *that* to link the cause and the event.

<**sentence id="S2"**><term id="T5" lex="mice" sem="Multi_cell">Mice</term> transgenic for the <term id="T6" lex="human_T_cell_leukemia_virus_(HTLV-I)_Tax_gene" sem="DNA_domain_or_region"><term id="T7" lex="human_T_cell_leukemia_virus" sem="Virus">human T cell leukemia virus</term> (<term id="T8" lex="HTLV-I" sem="Virus">HTLV-I</term>) <term id="T9" lex="Tax" sem="Protein_molecule">Tax</term> gene</term> develop <term id="T10" lex="fibroblastic_tumor" sem="Tissue">fibroblastic tumors</term> that express <term id="T11" lex="NF-kappa_B-inducible_early_gene" sem="DNA_family_or_group">NF-kappa B-inducible early genes</term>.**</sentence><event id="E1"**><type class="Cell_differentiation"/><theme idref="T10"/><clue>Mice transgenic for the human T cell leukemia virus (HTLV-I) Tax gene <clueType>develop</clueType> fibroblastic tumors that express NF-kappa B-inducible early genes.</clue>**</event><event id="E2"**><type class="Gene_expression"/><theme idref="T11"/><cause idref="T10"/><clue>Mice transgenic for the human T cell leukemia virus (HTLV-I) Tax gene develop fibroblastic tumors

<linkCause>that</linkCause> <clueType>express</clueType> NF-kappa B-inducible early genes.</clue></**event**>

(a)



(b)

Figure 8-2. A sentence and event annotation.

9. References

- Term annotation
  - Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi and Jun'ichi Tsujii. (2003). **GENIA corpus - a semantically annotated corpus for bio-textmining.** Bioinformatics. 19(suppl. 1). pp. i180-i182. Oxford University Press.
  - Ohta, Tomoko, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. (2002). **GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain.** In the Proceedings of the Human Language Technology Conference (HLT 2002).
- Part-of-speech annotation
  - Tateisi, Yuka and Jun'ichi Tsujii. (2004). **Part-of-Speech Annotation of Biology Research Abstracts.** In the Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004). IV. pp. 1267-1270.
  - Marcus, M. P., et al., **Building a Large Annotated Corpus of English: The Penn Treebank**, Computational Linguistics, Vol. 19, No. 2, 1994, pp. 313-330.
- Tree annotation
  - Tateisi, Yuka, Akane Yakushiji, Tomoko Ohta and Jun'ichi Tsujii . (2005). **Syntax Annotation for the GENIA corpus.** In the Proceedings of the IJCNLP 2005, Companion volume. pp. 222--227
  - Beis.A., Ferguson,M., Katz,K., and MacIntire,R.(1995). **Bracketing Guidelines for Treebank II Style: Penn Treebank Project**, University of Pennsylvania
- Event annotation
  - Alphonse, E., et al., **Event-based Information Extraction for the Biomedical Domain: the Caderige Project**, In the Proceedings of the COLING Workshop on Natural Language Processing in Biomedicine and its Applications, 2004, pp. 43–49.
- Coreference annotation
  - MEDCo project homepage: http://nlp.i2r.a-star.edu.sg/medco.html