TR-CS-UT-2006-5*

# GENIA Annotation Guidelines for Treebanking

Yuka Teteisi[2], Akane Yakushiji[3], Tomoko Ohta[1,] Jun'ichi Tsujii[1,4]

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

Date of Release: 7 November 2006



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

## About this document

This document describes the "local rule" of the GENIA treebank, as of GTB Beta 2 (500-abstract set) version. In this document, examples are shown shown in boxes and in bracketed form, not in XML, for visibility where each constituent is preceded by the label showing the "cat" attribute or the one showing the combination of "cat" and other attributes such as "syn" or "func". For example, the bracket notation

| (NP-COOD:(NP: John) and (NP: Mary)) |
|---|

corresponds to the following XML notation.

| <cons cat="NP" syn="COOD"><cons cat="NP">John</cons> and <cons cat="NP">MAary</cons></cons> |
|---|

A null constituent is marked by a symbol where the type of the constituent is surrounded by two asterisks in this document.

| (ADJP-COOD (ADJP: Ig-6-independent) and (ADJP (NP *QSTN*)-dependent)) |
|---|

For the actual XML encoding refer to "GENIA corpus manual -- Encoding schemes for the corpus and annotation" (TR-NLP-UT-2006-1).

## The scheme

We used The Penn Treebank II guidelines† with the following modifications.

1. When noun phrase consists with sequence of nouns, the internal structure is not necessarily shown. Some technical terms, not all, are bracketed, especially when they are parts of the coordinated phrases.

2. The phrases that should be labeled as NX and NAC in the constructions where parts of NP are coordinated are also labeled as NP, and we do not use NX and NAC tags at all.

| (NP:tasty (NP:(NP:California oranges) and (NP:Fuji apples))) |
|---|

3. Semantic tags used for the adverbials except TMP in the original PTB are left unused. Adverbials are just marked ADV.

| (S: (SBAR-ADV: For cells of the innate immune system to mount a host defence response to infection), (S: they must recognize products of microbial pathogens ...)) |
|---|

4. When single-word elements of the same syntactic category are coordinated, the coordination is not explicitly marked in the original PTB but was bracketed with flat structure instead. In the GENIA , single-word elements were also tagged.

| (NP-COOD:(NP:John) and (NP:Mary)) |
|---|

---

†Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. **Bracketing Guidelines for Treebank II Style: Penn Treebank Project**, University of Pennsylvania, 1995.

| |
|---|
| She (VP-COOD:(VP:smiled) and then (VP:cried)).<br>(VP:(VP-COOD:(VP:buy (NP *RNR*-1)) and (VP:eat (NP *RNR*-1))) (NP-1:oranges))<br>(NP:a (ADJP-COOD:(ADJP:beautiful), (ADJP:smart)) cat) |

The first three modifications are introduced in order to simplify the annotation process by linguists (non-biologists). Namely, we aim to annotate the structure determined without deep domain knowledge. On the other hand, the fourth one is based on the request of our colleagues who are working on HPSG grammar learning.
Although we were aware that a treebank corpus was being created as a part of in the Penn BioIE project‡, the guideline here is independent of theirs. Comparison of two schemes is one of the future works to be done. The problems not addressed in this document were handled using the original PTB manual and looking into similar cases in the WSJ corpus.

## Additional rules

With exception of the modifications mentioned above, we tried to follow the PTB manual as closely as possible, but due to the nature of scientific abstract there are some cases the original guideline did not fit. We treated them by case-by-case basis so that there may be inconsistent markups. The following cases are the ones that were generalized and established as rules.

1. In the case of the coordination structures where latter half of the phrases are ellipsed to form a structure like a mirror image of the "RNR" structure, no special tags or attributes were defined in the original PTB scheme. As of GTB Beta 2, we did not introduce a special value of the null attribute for that structure, but use QSTN (*?* in the original PTB) for the ellipsed constituent instead. As a result, the information of what is ellipsed in the structure is dropped, unlike RNR where the null constituent is co-indexed.

| |
|---|
| (NP: (NP: NF kappa B) (NP-COOD ((NP *QSTN*) protein) and ((NP *QSTN*) gene)) |

2. Case of coordination structures where a part of a word or words joined by a hyphen was coordinated. This is also handled by using QSTN as is in the following example.

| |
|---|
| (ADJP-COOD (ADJP: Ig-6-independent) and (ADJP (NP *QSTN*)-dependent)) |

‡ Colin Warner, Ann Bies, Christine Brisson, Justin Mott, **Addendum to the Penn Treebank II Style Bracketing Guidelines: BioMedical Treebank Annotation**, http://www.cis.upenn.edu/~bies/bioie/TBguidelines-addendum.pdf

This decision is in order to align the constituents of trees to token boundaries in the existing POS corpus. However, revision of tokenization criteria is in plan. When the tokenization criteria in the POS corpus are revised, the bracketing will also be changed.

3. The POS corpus should be used as a reference in case the annotator is not sure what category should be assigned. If the annotator is confident, (s)he can assign the category or categories contradicting the POS assigned to the head of the phrase.

4. Errors in the original text should be handled as if the structure was correct. When there is a duplication error, the same structure should be assigned to both of the duplications.

## Individual cases

This section includes the agreement on treatment of less generalized cases. The examples shown here can be used as references in case of uncertainties in the future, but the decisions here are more or less tentative in nature.

### Similar to~

"Similar to~" at the beginning of a sentence is tagged as ADVP, although it is a sequence of "adjective + prepositional phrase".

| |
|---|
| (S: (ADVP: Similar to LEF-1), (S: ALY can stimulate transcription in the context of the TCR alpha enhancer but apparently not when tethered to DNA through an heterologous DNA-binding domain) .) |

This is decided so because "regardless of ~" was tagged as ADVP in the original PTB manual (p.308) and two of the three occurrences of "contrary to~" were tagged as ADVP in the WSJ section of PTB.

### "For NOUN to VERB" construction indicating purposes

The construction is tagged as SBAR with ADV function, following the PTB manual.

| |
|---|
| (S: (SBAR-ADV: For cells of the innate immune system to mount a host defence response to infection), (S: they must recognize products of microbial pathogens ...)) |

### Numbers not in the prenominal position

Numbers which are not prenominal modifiers are tagged as NP as QP can only be used in the prenominal position according to the PTB manual.

| |
|---|
| (S: (NP: n) (VP: = (NP: 4)))<br>(ADJP: less (PP: than (NP: 0.07))) |

Numerals with units are treated as if the numeral modifies the unit. Thus, "about 50 %" is tagged as (NP: (QP: about 50) %).

**In vitro, etc.**

The modifiers "in vitro", "in vivo", "in situ", and "de novo" are tagged as ADJP if they are nominal modifiers and ADVP if they are verbal modifiers, and not tagged as PP.

| |
|---|
| (NP: (ADJP: in situ) hybridization) |
| (NP: (ADJP: In vitro-translated) proteins) |
| (VP: is (VP: regulated (NP: *NONE*) (ADVP: in vivo))) |

**Nominal labels**

Some abstracts have labels before the subparagraphs like OBJECTIVE：, METHOD：, RESULTS：, CONCLUSIONS：. Such labels are tagged as NP and regarded as a part of the first sentence, which is tagged as NP.

| |
|---|
| (NP: (NP: METHOD) ：(S: (NP-SBJ: They) (VP: measured the number of lymphocyte cytosolic glucocorticoid receptors))) |

**List numbers**

Numbers indicating list items are tagged as LST regardless of whether it is at the beginning of a sentence or in the middle of a sentence. If a list is inside a sentence, treat the list as coordination as an analogy to comma-separated coordination.

| |
|---|
| (S: (NP-SBJ: We) (VP: show (SBAR: that ：(S-COOD: (S: (LST: 1]) (S: ...)) (S: (LST: 2]) (S: ...) )))) |

**References**

Reference(s) at the end of a sentence is regarded as included in the sentence and treated as a modifier to the whole sentence.

| |
|---|
| (S: (S: We have identified and separated four nuclear proteins from C81-65-45 cells) (PRN: **(** (NP: Salahuddin et al., 1983) **)** ) ) |


## Acknowledgment

We are grateful to the annotators Emi Maekawa and Akiko Sagaguti.