Annotation Guidelines for Modifying Penn Treebank Trees into Binary Trees

MAEKAWA Emi, March 26, 2014

1. Basic Tasks

The primary task here is to convert Penn Treebank (hereafter, PTB) trees (automatically parsed trees based on the PTB bracketing and Par-of-Speech tagging guidelines) into binary trees. POS annotation errors also should be corrected if any as a result of automatic annotation.

2. Text to Annotate

2-1. Structure

In the PTB annotation scheme, POS and syntax annotation guidelines are written separately and in its syntax annotation, head elements are placed at the highest level of each layer.

In the following example, the word "love" is the head of the VP phrase “love Paris” and should not be given a VP label like in (VP (VP love) (NP Paris)).

```
(S (NP I)
   (VP love
       (NP Paris))).
```

Our scheme incorporates POS labels into syntax trees at the lowest level of each layer. Each individual morpheme (word) is given a POS label.

```
(S (NP <PRP I>)
   (VP <VBP love>
        (NP <NNP Paris>)))
<PERIOD .>
```

* Syntax and POS labels are shown by using ( ) and < > respectively in this manual. Parentheses other than ( ), like [ ] and { }, may be used for readers’ easier understanding.

2-2. Tag Set

Tags using symbols in the PTB tag set are modified as follows:

.	=>	PERIOD	,	=>	COMMA
:	=>	COLON	$	=>	DOLLAR
#	=>	SHARP	``	=>	DQL
"	=>	DQR	PRP$	=>	PRPD

"CD" tags that are given to numbers in the PTB tag set are not used in our tag set. Instead, numbers are labeled "NN".

3. N-ary Trees and Binary Trees

In Penn Treebank structure, one parent element can have any number of children, whereas in our structure, every parent element should consist of two or less children.

Example: `I went to school yesterday.`

In PTB structure, the VP phrase has three children as follows:

```
(S(S(NP <PRP I>)[VP <VBD went> (PP <TO to> (NP <NN school>)) (NP <NN yesterday> ])
<PERIOD.>))
```

In our scheme, any level that has three or more children should be modified to have two or less children (in the following, "went" and "to school" are grouped into one VP phrase):

```
((S(S(NP <PRP I>) [VP (VP <VBD went> (PP <TO to> (NP <NN school>))) (NP <NN yesterday>])
<PERIOD.>))
```

Example: `Founded by James Brown in 1891, ALMOT is an organization bringing a variety of unique program to New Yorkers.`

The following is the PTB bracketing results of the above sentence (POS labels are also shown by using ( )):

```
(S (S A1 (VP (VBN B1 Founded)
             (PP B2 (IN by)
                    (NP (NNP James)
                        (NNP Brown)))
             (PP B3 (IN in)
                    (NP (NN 1891)))))
   (COMMA A2 ,)
   (NP A3 (NNP ALMOT))
   (VP A4 (VBZ is)
          (NP (NP (DT an)
                  (NN organization))
```

```
                (VP (VBG C1 bringing)
                    (NP C2 (NP (DT a)
                               (NN variety))
                           (PP (IN of)
                               (NP (JJ unique)
                                   (NN programs))))
                     (PP C3 (TO to)
                            (NP (NNP New)
                                (NNPS Yorkers))))))
        (PERIOD A5 .))
```

In the above structure, elements marked with a red alphabet and number are the children of one element that should be modified to have two or less children.
The structure of the element that has A 1-5 children can be modified as follows:

```
(S (S Founded by James Brown in 1891)
    (COMMA ,)
    (NP ALMOT)
    (VP is an organization bringing a variety of unique programs to New Yorkers)
    (PERIOD .))
```

=> Divided into 2 elements : the period and the rest.

```
(S (S Founded by James Brown in 1891)
      (COMMA ,)
      (NP ALMOT)
      (VP is an organization bringing a variety of unique programs to New Yorkers))
   (PERIOD .))
```

=> Divided into three elements "founded by... 1891", a comma, and "ALMOT is ... New Yorkers" (If an element contains a comma, that element can have three children. We will discuss this later).

```
(S (S (S Founded by James Brown in 1891)
       (COMMA ,)
       (S (NP ALMOT)
           (VP is an organization bringing a variety of unique programs to New Yorkers)))
   (PERIOD .))
```

4. Exceptions - Cases That Do Not Need to Be Binary

The following structures can be ternary trees:

- Coordination structure (two or more elements joined by using periods, commas, colons/semicolons, and coordinating conjunctions)
- Phrases or clauses containing a comma or colon/semicolon
- Parenthetical elements (PRN)
- Quotes (QT)

BASENPs (noun phrases that have a flat internal structure) can have any number of children.

4-1. BASENP - Noun Phrases with Flat Internal Structure

Noun phrases that have no internal structure according to the PTB scheme can have any number of children. In our scheme, such flat NPs are labeled "BASENP" instead of "NP".

Example: `My sister Mary => (BASENP <PRPD My> <NN sister> <NNP Mary>)`
Wrong results: `(NP (BASENP <PRPD My> <NN sister>)(BASENP <NNP Mary>))`

4-2. Coordination Structures and Quotations

In principal, coordination structures, phrases containing commas or colons/semicolongs, quotations, and insert clauses can have three children:

- Coordination containing CC or CONJP

```
Tom and Jack
  => (BASENP <NNP Tom> <CC and> <NNP Jack>)
I bought and ate apples.
  => ... (VP (VP <VBD bought> <CC and> <VBDate>) (BASENP <NNS apples>))
```

- Elements containing a comma
  - Coordination structure:

```
red apples, oranges and bananas
```

```
(NP1 (NP2 (BASENP <JJ red>
                  <NNS apples>)
          <COMMA ,>
          (BASENP <NNS oranges>)
     <CC and>
     (BASENP <NNS bananas>)))
```

* Since NP1 has "and" as its child, it can have three children "red apples, oranges", "and", and "bananas", while NP2 has a comma therefore it can also have three children "red apples", ",", and "oranges". (The numbers following NP as "1" in "NP1" are used here to make readers to understand more easily. Actual tags do not have such numbers)

- Commas Used as Punctuation Marks

When a comma is used as a punctuation mark, the parent element can have three children.

```
Yesterday, he was there.
(S (S1 (BASENP <NN Yesterday>)
       <COMMA ,>
       (S (NP <PRP he>)
          (VP <VBD was>
              (ADVP <RB there>))))
   <PERIOD .>)
```

* S1 includes a "," therefore can have three children "Yesterday", ",", and "he was there".

- Colons and Semicolons

Colons and semicolons are treated the same as commas.

- Quotes Surrounded by DQL and DQR

```
He is called "BKB". => ... (QT <DQL "> (BASENP <NNP BKB>) <DQR">)
```

- Insert Phrases Using Commas or Parentheses

```
Bob, who is my brother, will come.
```

```
(S (S (NP (BASENP <NNP Bob>)
          (PRN <COMMA ,>
               (SBAR (WHNP <WP who>)
                     (S (VP <VBZ is>
                            (BASENP <PRPD my>
                                    <NN brother>))))
               <COMMA ,> ))
       (VP <MD will>
           (VP <VB come>)))
   <PERIOD .>)
```

5. Specifications for Binary Branching

This section describes cases that cannot be annotated using the PTB scheme, or cases that should be annotated differently from PTB trees.

5-1. Head Elements

In PTB trees, head elements are basically placed directly under the bracket they belong to.

```
I know and like him.
(S (BASENP I) (VP know and like (BASENP him))).
```
* POS labels are omitted for easier understanding of structure.

The above bracketing shows that "know and like" is the head of the VP and "him" is the modifier placed at the lower level. In our binary branching, the head "know and like" is also to be bracketed as a VP so that the higher level VP can have two children. This is the biggest difference from the basic PTB structure.

```
(S (BASENP I) (VP (VP know and like) (BASENP him))).
```
*POS labels omitted

```
(S (S (BASENP <PRP I>)
      (VP (VP <VBP know>
              <CC and>
              <VBP like>)
          (BASENP <PRP him>))
   <PERIOD .>)
```

In addition, elements that should be NXs do not need to be given NX labels.

5-2. Coordination Structures

5-2-1. Coordinating Conjunctions

The following are conjunctions that make coordinates

POS labels: CC　COMMA　COLON
Syntax labels: CONJP

In addition to CONJP specified in the PTB guidelines, the combinations of a comma and a CC, and a colon/semicolon and a CC are bracketed CONJP.

```
(BASENP <NNS oranges> (CONJP <COMMA ,> <CC or>) <NNS apples>)
```

5-2-2. Coordinates Consisting of Two Elements

When the coordinated elements in a coordinate are linked by a coordinating conjunction (CC, CONJP), comma, or colon/semicolon, that coordinate can be a ternary tree.

```
apples and oranges
(BASENP <NNS apples>
        <CC and>
        <NNS oranges>)

apples rather than orangges
(BASENP <NNS apples>
        (CONJP <IN rather>
               <IN than>)
        <NNS oranges>)
```

Verb phrases are to be analyzed likewise:

```
The machine inputs and outputs information.
```

PTB bracketing of the VP level:

```
(VP <VBZ inputs> <CC and> <VBZ outputs> (BASENP <NN information>))
```

Our bracketing:

```
(VP (VP <VBZ inputs> <CC and> <VBZ outputs>) (BASENP <NN information>))
```

"inputs and outputs" is the head of the highest VP but can be grouped as a VP. It can have three children because it is a coordinate.

S clauses and phrases:

```
Tom likes dogs but Mike likes cats.
(S (S Tom likes dogs) <CC but> (S Mike likes cats)).

(S (S (S (BASENP <NNP Tom>)
         (VP <VBZ likes>
             (BASENP <NNS dogs>)))
      <CC but>
      (S (BASENP <NNP Mike>)
          (VP <VBZ likes>
              (BASENP <NNS cats>))))
   <PERIOD .>)
```

5-2-3. Coordinates Consisting of Three or More Elements

If there are three or more coordinated elements in a coordinate, the coordinated elements are to be grouped in the appearing order.

```
[NP {NP 1,2,3} {CC and} {BASENP 4}]
=>[NP{NP [NP 1,2] [COMMA,] [BASENP 3]} {CC and} {BASENP 4}]
=>[NP{NP (NP <BASENP 1> <COMMA,> <BASENP 2>) (COMMA,) (BASENP 3)} {CC and} {BASENP 4}]

  3 cats, 2 dogs, and other pets.
  => (NP (NP 3 cats, 2 dogs) (CONJP , and) (BASENP other pets))
  => (NP (NP (BASENP 3 cats)
             <COMMA ,>
             (BASENP 2 dogs))
         (CONJP , and)
         (BASENP other pets))    * POS labels and lower structures omitted.
```

Exceptions: Titles are to be grouped from the last element since they are usually constructed as [main headline (middle headline <subheadline>)]. (Note that our POS tagging scheme for titles is different from the PTB's. Details will be described later.)

```
Special Issue : Preoperative diagnosis : Lang cancer.
(NP (BASENP Special Issue)<COLON :>(NP Preoperative diagnosis : Lung cancer.)):

(NP (BASENP Special Issue)
    <COLON :>
    (NP (BASENP Preoperative diagnosis)
        <COLON :>
        (NP (BASENP Lung cancer)
            <PERIOD .>)))
```

5-3. Subordinate Clauses (SBAR)

5-3-1. Basic Structure

In the PTB scheme, conditional, temporal, or other such SBARs are attached under VP if they follow the main clause, while our scheme places such SBARs outside the main clause as follows:

```
He was there when I came home.
```

PTB bracketing:

```
(S (S (BASENP He) (VP was (ADVP there) (SBAR when I came home))).)
(S (S (BASENP He)
      (VP was
            (ADVP there)
            (SBAR when I came home))).)
```

* POS labels and the lower level structure of SBAR omitted.

Our bracketing

```
(S (S (S (BASENP He) (VP was (ADVP there)) (SBAR when I came home))).)
(S (S (S (BASENP <PRP He>)
         (VP (VBD was>
              (ADVP <RB there>))
      (SBAR (WHADVP <WRB when>)
            (S (BASENP <PRP I>)
               (VP <VBD came>
                    (ADVP <RB home>))))))
   <PERIOD .>)
```

The same bracketing is given even if the subject in an SBAR clause is not explicitly written.

```
He eats his dinner while watching television.
(S (S (S (BASENP He) (VP eats his dinner)) (SBAR while watching television)).)
```

* Lower level structures omitted

However, S clauses such as "to infinitive" clauses expressing a purpose or reason and "-ing clauses" are labeled S and attached at VP level.

```
He works hard to pay for school.
(S (S (BASENP He) (VP (VP works (ADVP hard)) (S (VP to pay for school)))).)
```

```
He eats his dinner, watching television.
(S (S (BASENP He) (VP (VP eats (BASENP his dinner)), (S watching television))).)
```
* Lower level structures omitted

5-3-2. According to ...

When the phrase "according to" appears after the main clause (verb), the PP phrase governed by "according to" should not be attached to the VP phrase.

```
The president left for Paris, according to the spokesman.
(S (S (S The president left for Paris), (PP according to the spokesman)).)
```
* Lower level structures omitted

Note that "according to" modifying the preceding noun, often seen in patent documents, is attached to the noun phrase.

```
The machine 21 includes the device 24 according to the claim 1.
The machine 21 includes (NP (BASENP the device 24) (PP according to the claim 1)).
```
* Lower level and irrelevant structures omitted

5-3-3. Coordinated Verb Phrases Sharing the Subject

When two or more VPs share the subject but do not share SBAR (e. g. He went out but came back soon because it was raining outside.), coordination is not at the VP level but at the higher S level, and the VP lacking the explicit subject is left without the subject.

```
  (S (S (S He went out) but (S came back soon because it was raining outside)).)
=>(S(S(S He went out) but (S (S came back soon)(SBAR because it was raining outside))).)

  (S (S (S He went out)
        <CC but>
        (S (S came back soon)
            (SBAR because it was raining outside))).)
```
* Lower level structures omitted

5-3-4. Relative Clauses

When a relative clause has a subordinate SBAR clause, the subordinate SBAR and the relative clause excluding that subordinate SBAR are placed at the same level as siblings.

```
This is the house that my father built when he was twenty.
```

* In the following analysis, only the NP phrase "the house ... twenty" is shown for easier understanding.

PTB bracketing (lower level bracketing omitted):

```
(NP (BASENP the house) (SBAR (WHNP that) (S (BASENP my father) (VP built (SBAR when he was twenty)))))
(NP (BASENP the house)
    (SBAR (WHNP that)
          (S (BASENP my father)
             (VP built
                 (SBAR when he was twenty)))))
```

Our bracketing:

```
(NP (BASENP the house) (SBAR (SBAR that my father built) (SBAR when he was twenty)))

(NP (BASENP <DT the>
            <NN house>)
    (SBAR (SBAR (WHNP <WDT that>)
                (S (BASENP <PRPD my>
                           <NN father>)
                     (VP <VBD built>)))
          (SBAR (WHADVP <WRB when>)
                (S (BASENP <PRP he>)
                   (VP <VBD was>
                       (BASENP <NN twenty>)))))))
```

Non-restrictive relative clauses are not attached to the main clause (verb) but are placed at the same level as the main clause S.

```
He said he had seen a UFO, which we all know is a lie.
(S (S (S He said he had seen a UFO), (SBAR which we all know is a lie)).)
```

* Lower level structures omitted

5-4. Parenthetical Phrases

Parenthetical phrases (PRN) are:

- Phrases or clauses that are surrounded by a pair of commas, colons, semicolons, or dashes.
- Phrases or clauses that are surrounded by parentheses.

Based on the above definition, appositive noun phrases that are set off by two commas are labeled PRN.

```
Tanaka, the prime minister, will be here.
(NP (BASENP <NNP Tanaka> )
    (PRN <COMMA ,>
         (BASENP <DT the>
                 <JJ prime>
                 <NN minister>)
         <COMMA ,>))
```

Other examples of parenthetical phrases:

Relative clauses:

```
Tokyo, where he was born, is the capital of Japan.
(NP (BASENP <NNP Tokyo>) (PRN <COMMA ,> (SBAR where he was born) <COMMA ,>) )
```

* The lower structure of SBAR omitted.

Brackets:

```
Association of South-East Asian Nations -LRB- ASEAN -RRB-
(NP (NP (BASENP <NNP Association>)
         (PP <IN of>
              (BASENP <NNP South-East>
                      <NNP Asian>
                      <NNPS Nations>)))
    (PRN <LRB -LRB->
          (BASENP <NNP ASEAN>)
          <RRB -RRB->))
```

5-5. Quotations.

An element starting with a DQL (opening quotation mark) and ending with a DQR (closing quotation mark) is labeled "QT", and the quoted part is labeled "QTC" if it has an end punctuation such as "." and ",", or labeled whatever is appropriate if it has no end punctuation.

```
He said "I will do that," last night.
(S(S(BASENP He) (VP(VP said (QT"(QTC (S I will do that),)")) (BASENP last night))).)
```

* POS and the structure of the quotation ommited.

```
That dog is called "Snoopy".
(S (S (BASENP That dog) (VP is (VP called (S (QT "(BASENNP Snoopy)"))))).)

(S (S (BASENP <DT That>
              <NN dog>)
      (VP <VBZ is>
          (VP <VBN called>
              (S (QT <DQL">
                     (BASENNP <NNP Snoopy>)
                     <DQR ">))))))
  <PERIOD .>)
```

If a quotation is divided into two sentences for any reason, the beginning part is labeled QTL and the ending part QTR.

```
He said that "This is intorelable. <=  the first sentence
I cannot stand it. " <= the second sentence

(S (BASENP He)(VP said (QTL " (S This is intorelable.))))
(QTR (S I can not stand it.) ")
```

5-6. Noun Phrases

5-6-1. BASENP

Noun phrases whose internal structure is to be left flat according to the PTB scheme are labeled BASENP (not NP) and do not have to have binary structure.

```
(BASENP <DT A> <JJ beautiful> <NN house>)
```

According to the PTB scheme, a phrase consisting of a noun and its appositive noun is labeled BASENP unless the appositive is preceded by a comma or article.

```
(BASENP <PRPD His> <JJ favorite> <NN girl> <NNP Tomoko>)
```

However, if the appositive is preceded by a comma or article, the noun and its appositive are separated into different noun phrases.

```
(NP (BASENP <PRPD His> <JJ favorite> <NN girl>) <COMMA ,> (BASENP <NNP Tomoko>))
```

■ Modifying Adjectives

When a noun is premodified by two single-morpheme adjectives linked with a CC, the whole noun phrase is labeled BASENP and the premodifier, consisting of two coordinating adjectives, is labeled ADJP.

```
a red and sweet apple
(BASENP a (ADJP red and sweet) apple)
=> (BASENP <DT a> (ADJP <JJ red> <CC and> <JJ sweet>) <NN apple>)
```

When the two adjectives are linked not with a CC but a comma, the whole noun phrase is NOT labeled BASENP but NP.

```
a red, sweet apple
(NP <DT a>
    (NP (ADJP <JJ red>
              <COMMA ,>
              <JJ sweet>)
        <NN apple>))
```

When an adjective is modified by a single-morpheme modifier (e.g. an adverb), the entire adjective phrase is labeled ADJP.

```
a very sweet wine
 => (BASENP <DT a> (ADJP <RB very> <JJ sweet>) <NN wine>)
the ink transporting carrier
 => (BASENP <DT the> (ADJP <NN ink> <VBG transporting>) <NN carrier>)
```

When an adjective is modified by a multi-morpheme modifier, the entire noun phrase is not labeled BASENP but NP.

```
a red and very sweet apple :
(NP a (NP (ADJP (ADJP red) and (ADJP very sweet)) apple))

(NP <DT a>
    (NP (ADJP (ADJP <JJ red>)
              <CC and>
              (ADJP <RB very>
                    <JJ sweet>))
        <NN apple>))

the red ink transporting carrier:
(NP the (NP (ADJP (BASENP red ink) transporting) carrier))

(NP <DT the>
    (NP (ADJP (BASENP <JJ red>
                      <NN ink>)
              <VBG transporting>)
        <NN carrier>))
```

c.f. (BASENP the red ink transporting technique) *Detailes will be described later.

When three or more coordinated adjectives premodify a noun, the entire noun phrase is not labeled BASENP but NP.

```
a red, sweet and fresh apple
```

```
(NP a (NP (ADJP (ADJP (ADJP red), (ADJP sweet)) and (ADJP fresh)) apple))

(NP <DT a>
    (NP (ADJP (ADJP (ADJP red)
                    <COMMA ,>
                    (ADJP sweet))
              <CC and>
              (ADJP fresh))
        <NN apple>))
```

■ Coordinated BASENP

When two single-morpheme nouns are coordinated, they make up a BASENP.

```
(BASENP <NNS apples> <CC and> <NNS oranges>)
```

When at least one noun phrase in a coordinate structure consists of two or more morphemes, each coordinated noun is labeled BASENP.

```
(NP (BASENP <JJ red> <NNS apples>) <CC and> (BASENP <JJ orange> <NNS oranges>))
```

* Note that "delicious apples and oranges" is labeled BASENP and has no internal structure.

When there are three or more coordinated single-morpheme nouns, each noun (noun phrase) is labeled BASENP.

```
apples, oranges and bananas
(NP (NP (BASENP <NNS apples>)
        <COMMA ,>
        (BASENP <NNS oranges>))
    <CC and>
    (BASENP <NNS bananas>))
```

5-6-2. Ing-form Adjectives

Basically, ing-form verbs are labeled according to the PTB scheme. The following are the frequently

appearing patterns and their labeling based on our scheme.

・[the object of -ing] + [-ing] + [subject of -ing]:

The POS for the -ing element is VBG. The object and -ing make one ADJP phrase

```
vitamin transporting substance
* meaning : substance (subject) that transports vitamin (= object)
(BASENAP (ADJP <NN vitamin> <VBG transporting>) <NN substance>)
```

・[the object of -ing] + [-ing] + [noun (not the subject of -ing)]:

The POS for the -ing element is NN, not VBG, and it does not make an ADJP phrase

```
vitamin transporting method
* meaning : method for transporting vitamin (object)
(BASENP <NN vitamin> <NN transporting> <NN method>)
```

* The PTB scheme says that the POS for the above type -ing is (NN|VBG), but our scheme does not use "|".

・When -ing and the object of -ing are hyphenated:

The POS is always JJ.

```
vitamin-transporting substance => (BASENP <JJ vitamin-transporting> <NN substance>)
vitamin-transporting method => (BASENP <JJ vitamin-transporting> <NN method>)
```

5-6-3. NAC

According to the PTB scheme, NAC phrases are bracketed as follows (POS labels omitted):

```
Secretary of State Johnson => (NP (NAC Secretary (PP of (NP State))) Johnson)
```

In our scheme, unlike the PTB's, head nouns in NAC structure are labeled NP or BASENP.

```
(NP (NAC (BASENP Secretary) (PP of (BASENP State))) Johnson)
```

With POS labeles (complete annotation):

```
(NP (NAC (BASENP <NNP Secretary>)
         (PP <IN of>
             (BASENP <NNP State>)))
```

```
<NNP Johnson>)
```

Since our scheme treats NAC phrases and other noun phrases the same way, the "NAC" labels are not necessarily used. If NAC has already been given as a result of automatic tagging, leave it as it is. Otherwise, annotators can substitute NP for NAC.

5-6-4. QP (QUANTIFIER PHRASE)

Noun phrases with a QP phrase having a preposition(s) at its lower tier(s) are labeled not BASENP but NP and bracketed as follows:

■ *more than / less than*

"more than" is labeled ADVP. The POS of "more" is RBR.

```
more than 5 million cars
(NP (QP (ADVP <RBR more> <IN than>) (BASENP <NN 5> <NN million>)) <NNS cars>)
```

■ *at least*

"at least" is labeled ADVP. The POS of "least" is JJS.

```
at least ten black cats
(NP (QP (ADVP <IN at> <JJS least>) <NN ten>) (BASENP <JJ black> <NNS cats>))
```

■ *in, out of,* and others

Not QP unless it modifies the succeeding noun.

```
One in four agreed to the plan.
=> (NP (BASENP <NN One>) (PP <IN in> (BASENP <NN four>)))...
```

QP when it modifies the succeeding noun.

```
One in four persons agreed to the plan.
=> (NP (QP (BASENP <NN One>) (PP <IN in> (BASENP <NN four>))) <NNS persons>)...
```

■ *between*

Prepositional phrases including "between" are labeled QP and the POS of "between" is RB.

```
between 5 and 10 minutes
(NP (QP <RB between> (BASENP <NN 5> <CC and> <NN 10>)) <NNS minutes>)
```

■ *up to*

"up to" is labeled ADVP. The POS of "up" is RB.

```
up to 5 million cars
(NP (QP (ADVP <RB up> <TO to>) (BASENP <NN 5> <NN million>)) <NNS cars>)
```

■ *as many as / as much as / as [adjective] as*

"as [adjective] as" is labeled ADJP, with its internal structure appropriately bracketed, and makes QP with the succeeding numerical expression. The POS of "as" preceding the adjective is RB.

```
as many as 5 million cars
(NP (QP (ADJP (ADJP <RB as>
                    <JJ many>)
              <IN as>)
         (BASENP <NN 5>
                 <NN million>))
    <NNS cars>)
```

5-6-5. Proper Nouns

According to the PTB scheme, the POS's of the capitalized words in a name are either NNP or NNPS, and if the name has the lower level tier(s), it is syntactically bracketed.

```
Gone with the Wind
(S (VP Gone (PP with (BASENP the Wind)))) * Without POS labels
(S (VP <NNP Gone> (PP <IN with> (BASENP <DT the> <NNP Wind>)))) * With POS labels

the United States of America
(NP (BASENP the United States) (PP of (BASENP America)))
(NP (BASENP <DT the> <NNP United> <NNPS States>)(PP <IN of> (BASENP <NNP America>)))
```

In our annotation, unlike the PTB POS tagging, the titles of intellectual works (e.g. literary works, academic papers, laws, songs, art works, movies etc.) are not labeled NNP or NNPS. For the POS of other types of proper names (e.g. names of places, persons, organizations, products etc.), follow the PTB POS tagging guidelines. Based on these rules, "Gone with the Wind" is bracketed as follows (the United States of America is bracketed according to the PTB scheme):

```
(S (VP Gone (PP with (BASENP the Wind)))) * Without POS labels
(S (VP <VBN Gone> (PP <IN with> (BASENP <DT the> <NN Wind>)))) * With POS labels
```

5-6-6. Titles

As mentioned above, titles of intellectual works are not labeled NNP or NNPS. However, since the names of newspapers and magazines are often difficult to distinguish from their publishers' names, they are labeled NNP/NNPS. The following are notes when annotating titles with non-flat structure.

■ Titles with Non-flat Structure

When a title has a nested structure consisting of a heading and subheadings, the lowest-level subheadings are first to be bracketed together.

```
Special Issue : Preoperative diagnosis : Lang cancer
(NP (BASENP Special Issue) <COLON :> (NP Preoperative diagnosis : Lung cancer))

(NP (BASENP Special Issue)
    <COLON :>
    (NP (BASENP Preoperative diagnosis)
        <COLON :>
        (BASENP Lung cancer)))
```

In the case of "[Heading 1] [colon] [Heading 2] [period]", the period is grouped with [Heading 2].

```
Special Issue : Preoperative Diagnosis.
(NP (BASENP Special Issue) <COLON :> (NP Preoperative Diagnosis.))
=> (NP (BASENP Special Issue)
   <COLON:>
   (NP (BASENP Preoperative Diagnosis)
       <PERIOD.>))
```

■ Non-nominal Titles

When a title is not a noun phrase, it does not have to be a noun but is given an appropriate structure (for example, it can be an S).

```
Gone with the Wind → (S (VP Gone (PP with (BASENP the Wind)))) * POS labels omitted
```

When a heading and a subheading(s) belong to different syntactic categories (e.g. a noun phrase and a sentence), the phrase node that group them is labeled UCP.

```
Special Issue : Our Future : Where Are We Going?
(UCP (BASENP Special Issue)
     <COLON :>
     (UCP (BASENP Our Future)
          <COLON :>
          (SBARQ Where Are We Going?  ))) * Lower level structures omitted.
```

5-6-7. Chemical Substance Names and Formulas

Symbols appearing in substance names such as "+", "-", and "[ ]" are not given any syntactic consideration but treated like nouns. Unless a substance name contains prepositional phrases, it is labeled BASENP. The POS's of "[" and "]" are LRB and RRB respectively, and the POS's of other types of symbols are SYM.

```
[C3H3O]+ ions
(BASENP <LRB [>
        <NN C3H3O>
        <RRB ]>
        <SYM +>
        <NNS ions>)

2 - dimethoxy'- binaphthyl' - acid derivatives
(BASENP <NN 2>
        <SYM ->
        <NN dimethoxy>
        <SYM ’>
        <SYM ->
        <NN binaphthyl>
        <SYM ’>
        <SYM ->
```

<NN acid>
<NNS derivatives>)

5-7. VERB PHRASES

5-7-1. Negation: not

Unlike the PTB scheme, the negative element "not" is not left unlabeled but grouped with the verb or auxiliary verb.

```
He can not swim.
(S (S (BASENP He) (VP (VP can not) (VP swim))).)

(S (S (BASENP <PRP He>)
      (VP (VP <MD can>
              <RB not>)
          (VP <VB swim>)))
   <PERIOD .>)
```

Even when "not" is followed by an adjective (e. g. This is not beautiful.), the "not" is not grouped with the adjective but with the verb (or auxiliary verb).

```
This is not beautiful.
(S (S (BASENP This) (VP (VP is not) (ADJP beautiful))).)
```
* POS labels omitted.

However, if the "not" has to be grouped with the adjective for a structural reason, the not is grouped with the adjective (or other appropriate elements).

```
She is not smart but pretty.
```
("pretty" and "not smart" should be coordinated)
```
(S (S (BASENP She) (VP is (ADJP (ADJP not smart) but (ADJP pretty)))).)

(S (S (BASENP <PRP She>)
      (VP <VBZ is>
          (ADJP (ADJP <RB not>
                      <JJ smart>)
                <CC but>
                (ADJP <JJ pretty>))))
   <PERIOD .>)
```

5-7-2. Position of ADVP

When an adverbial phrase (ADVP) is inserted in a verb phrase consisting of two or more verbs/auxiliary verbs (e.g. have been), the ADVP is grouped with the verb/auxiliary verb succeeding the ADVP.

```
I have always wanted to see this film.
(S (S (BASENP I)
      (VP have always wanted to see this film)). )

=> (S (S (BASENP I)
         (VP <VBP have>
             (VP always wanted to see this film))). )

=> (S (S (BASENP I)
         (VP <VBP have>
             (VP (ADVP always)
                 (VP wanted to see this film)))). )

=> (S (S (BASENP I)
         (VP <VBP have>
             (VP (ADVP always)
                 (VP <VBN wanted>
                     (S to see this film))))). )
```

* POS labels and lower level structure omitted

5-7-3. Verb Phrases with Multiple Complements and Modifiers.

When a verb has multiple children (complements or modifiers), each child is bracketed with the verb, not with its siblings.

`I saw him yesterday here.` <= "him", "yesterday", and "here" depend on the verb "saw".

PTB tree:

```
(S (BASENP I) (VP saw (BASENP him) (BASENP yesterday) (ADVP here))).
```

Our tree:

```
(S (S (BASENP I) (VP (VP (VP saw him) (BASENP yesterday)) (ADVP here))).)
```

```
(S (S (BASENP I)
       (VP (VP (VP <VBD saw>
                   (BASENP him))
               (BASENP yesterday))
           (ADVP here)))
   <PERIOD .>)
```

* lower level structures omitted.

5-8. It-Clefts and It-Extraposition

■ It-Clefts

According to the PTB scheme, the sentence "It is him who broke the window." is bracketed as follows:

```
(S (BASENP It)
   (VP is
       (BASENP him)
       (SBAR (WHNP who)
             (S (VP broke
                    (BASENP the window))))).)
```

In our binary branching, the verb "is" and "him" are first to be grouped (not "him" and "who").

```
(S (BASENP It)
   (VP (VP is (BASENP him))
       (SBAR (WHNP who) (S (VP broke (BASENP the window))))).)
```

* POS labels omitted.

■ It-Extraposition

In it-extraposition, the verb (be) and the predicate immediately following the verb are grouped first, and the clause or phrase that replaces "it" is placed outside this [verb] + [predicate] bracket.

```
It is desirable to fix the handle.
(S (S (BASENP It) (VP (VP is desirable) (S to fix the handle))).)

It's good that you are here.
(S (S (BASENP It) (VP (VP 's good) (SBAR that you are here))).)

I find it annoying that they make noise.
(S (S (BASENP I) (VP find (S (S it annoying) (SBAR that they make noise)))).)
```

* POS labels and lower level structures omitted.

5-9. SINV (Inversion)

According to the PTB scheme, inverted sentences are bracketed as follows:

```
Never hadI seen such a place.
=> (SINV (ADVP Never) had (NP I) (VP seen such a place).)
```

In our scheme, inverted sentences are bracketed as follows:

```
(SINV (ADVP Never) had (BASENP I) (VP seen such a place).)

=> (SINV (SINV (ADVP Never) had (BASENP I) (VP seen such a place)) <PERIOD .>)

=> (SINV (SINV (ADVP Never) (SINV had (NP I) (VP seen such a place))) <PERIOD .>)

=> (SINV (SINV (ADVP Never)(SINV had (S (NP I) (VP seen such a place)))) <PERIOD .>)

(SINV (SINV (ADVP <RB Never>)
            (SINV <VBD had>
                  (S (NP <PRP I>)
                      (VP <VBN seen>
                           (BASENP <PDT such>
                                   <DT a>
                                   <NN place.)))))
      <PERIOD .>)


Most surprising was her face.

PTB tree:
(SINV (ADJP Most surprising) (VP was) (BASENP her face).)
Our tree :
(SINV (SINV (ADJP Most surprising)
            (SINV (VP was)
                  (BASENP her face))).)
```

* POS labels omitted.

5-10. Questions (SQ and SBARQ)

■ SQ: Yes/No Questions

According to the PTB scheme, yes/no questions are bracketed as follows:

```
Do you know him?  => (SQ Do (BASENP you) (VP know (BASENP him))? )
```

According to our scheme, the portion excluding the auxiliary verb at the sentence beginning makes an "SQ". The auxiliary verb is not bracketed "VP" unless it is a multi-word phrase.

```
(SQ (SQ Do (SQ (BASENP you) (VP know (BASENP him))))? )
```

With POS labels:

```
(SQ (SQ <VBP Do>
        (SQ (BASENP <PRP you>)
            (VP <VB know>
                (BASENP <PRP him>))))
   <PERIOD ?>)
```

```
Is this your car?
PTB tree: (SQ Is (BASENP this) (BASENP your car)?)
Our tree: (SQ (SQ Is (SQ (BASENP this) (BASENP your car)))? )
  (SQ (SQ <VBZ Is>
          (SQ (BASENP this)
              (BASENP your car)))
      <PERIOD ?>)
```

Questions without an auxiliary verb is simply labeled S, not SQ.

■ SBARQ: WH Questions

```
What did you eat?
(SBARQ (SBARQ (WHNP What) (SQ did (SQ (BASENP you) (VP eat))))? )
```

```
(SBARQ (SBARQ (WHNP <WP What>)
              (SQ <VBD did>
                  (SQ (BASENP <PRP you>)
                      (VP <VB eat>))))
       <PERIOD ? >)
```

```
Who did this?
(SBARQ (SBARQ (WHNP Who) (SQ (VP did (BASENP this))))?)
```

```
(SBARQ (SBARQ (WHNP <WP Who>)
              (SQ (VP <VBD did>
                      (BASENP <DT this>))))
       <PERIOD ?  >)
```

```
Who are you? : "are" is not bracketed VP.
(SBARQ (SBARQ (WHNP Who) (SQ are (BASENP you))) ?)
```

```
(SBARQ (SBARQ (WHNP <WP Who>)
              (SQ <VBP are>
                  (BASENP <PRP you>)))
       <PERIOD ?>)
```

5-11. Miscellaneous

5-11-1. From A to B

When a combination of prepositional phrases "from ..." and "to ..." indicate the start point and end point respectively, they do not make a PP coordination. Each prepositional phrase depends on the main verb and makes VP with the main verb.

```
The device sends data from the machine A to the machine B.
```

* Only the relevant part (VP) is shown in the following example.

```
(VP (VP sends data from the machine A) (PP to the machine B)

=> (VP (VP (VP sends data) (PP from the machine A)) (PP to the machine B)))

=> (VP (VP (VP <VBZ sends>
                (BASENP <NNS data>))
           (PP <IN from>
               (BASENP <DT the>
                       <NN machine>
                       <NN A>)))
       (PP <TO to>
            (BASENP <DT the>
                    <NN machine>
                    <NN B>))))
```

When "from ..." and "to ..." indicate a range, they make a PP coordination.

```
Anybody from children to elder people can join us.
(NP (BASENP Anybody) (PP (PP from children) (PP to elder people)))

(NP (BASENP <NN Anybody>)
    (PP (PP <IN from>
            (BASENP <NNS children>))
        (PP <TO to>
            (BASENP <JJ elder>
                    <NNS people>))))
```

5-11-2. So-That Clauses

So-that clauses are labeled SBAR, placed outside the main VP as the sibling of the main S. The POS labels of both "so" and "that" are IN and they are bracketed as a single SBAR.

```
Turn off the radio so that I can sleep.
(S (S (S Turn off the radio) (SBAR (SBAR so that) (S we can sleep)).)

(S (S (S (VP (VP <VB Turn>
                 (PRT <RP off>))
             (BASENP <DT the>
                     <NN radio>)))
       (SBAR (SBAR <IN so>
                   <IN that>)
             (S (BASENP <PRP we>)
                (VP <MD can>
                    (VP <VB sleep>)))))
    <PERIOD .>)
```

* Note: The POS labels of "such" and "that" in subordinating conjunction are JJ and IN respectively.

5-11-3. Including ...

The POS of "including" that behaves like a preposition is VBG and the phrase governed by "including" is labeled PP.

```
(NP (BASENP <NN Everybody>) (PP <VBG including> (BASENP <NNP Tom>)))
```

In patent documents, "including" is often used as an ing-form verb meaning "containing" or "consisting of". In such cases, phrases governed by "including" are labeled VP, not PP.

```
The host 200 is a mainframe computer including a CPU 201, a memory 202, and a
storage medium 204.
```

```
(S (S (BASENP The host 200)
      (VP <VBZ is>
          (NP (BASENP a mainframe computer)
               (VP <VBG including>
                    (NP (NP (BASENP a CPU 201)
                            <COMMA ,>
                            (BASENP a memory 202))
                        (CONJP <COMMA ,>
                               <CC and>)
                        (BASENP a storage medium 204))))))
   <PERIOD .>)
```

* The POS labels of the elements under BASENP are omitted for easier understanding.

5-11-4. Both A and B / Either A or B

The phrases "both A and B" and "either A or B" are bracketed as follows.

```
Both Tom and Mike
(NP <CC Both> (BASENP <NNP Tom> <CC and> <NNP Mike>))

Either go out or come in.
(S (S (VP <CC Either> (VP go out or come in))).)

=> (S (S (VP <CC Either> (VP (VP go out) <CC or> (VP come in)))).)

=> (S (S (VP <CC Either>
             (VP (VP <VB go>
                     (PRT <RP out>))
                 <CC or>
                 (VP <VB come>
                     (PRT <RP in>))))))
       <PERIOD .>)
```