Addendum to the Penn Treebank II Style Bracketing Guidelines: BioMedical Treebank Annotation

Colin Warner, Ann Bies, Christine Brisson, Justin Mott

University of Pennsylvania
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104, USA

November, 2004

Contents

Introduction

This *Addendum* is meant to be used alongside *Bracketing Guidelines for Treebank II Style* (1995), as it contains the additions and changes to treebank annotation policy that were developed by the treebank annotation group as part of the ITR/E Biomedical Information Extraction project (http://bioie.ldc.upenn.edu, funded via award EIA-0205448 from the National Science Foundation's Information Technology Research (ITR) program). The treebank annotation for this project is primarily based on the Penn Treebank II guidelines (*Bracketing Guidelines for Treebank II Style, Penn Treebank Project*, Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. University of Pennsylvania Computer and Information Science Department Technical Report MS-CIS-95-06, LINC LAB 281, 1995).

The Penn Treebank II guidelines were followed as closely as possible, but the nature of the biomedical corpus has made some changes necessary or desirable. We have also taken this opportunity to address several long-standing issues with the original set of guidelines, with regard to NP structure in particular. This has resulted in the introduction of one new node label for sub-NP nominal substrings (NML), and the elimination of one node label (NX). NML, among other functions, replaces entirely the previous use of both NX and NAC *within NPs*. (NAC continues in its external-to-NP role.) The use of these node labels is described in the appropriate sections below. One additional empty category (*P*) has been added and is described in the appropriate sections below. *P* was introduced in order to improve the match-up of entity categories with treebank nodes. It is used as a placeholder to represent distributed modification in nominals and does not represent the trace of movement.

There has been a change in the formatting of indices: all indices are now on node labels, and index chains can be viewed as equivalence classes, eliminating the need for cascading index chains. Gapping indices are now completely independent of other indices, and are shown with "=" on the node labels of all gapping constituents (both in the template and in the conjuncts).

Note that the biomedical project (along with other current treebank projects) has adopted several changes in word-level tokenization, and this leads to a number of part-of-speech and structural differences as well. Many hyphenated words are now treated as separate tokens ("New York – based" would be four tokens, for example). These hyphens now have the part-of-speech tag HYPH. If the separated prefix is a morphological unit that does not exist as a free-standing word, it has the part-of-speech tag AFX. With chemical names and scientific notation in the biomedical corpus in particular, spaces and punctuation may occur within a single "token," which will have a single POS tag. The treebank annotation treats any token with a POS tag as a "word" for the purposes of single- or multi-word decisions as they affect the tree structure.

Please contact us with any questions you may have about biomedical treebank annotation!

Ann Bies
bies@ldc.upenn.edu
Linguistic Data Consortium
University of Pennsylvania
November, 2004

1 Changes in the Bracketing of Nominals for the BioMedical Information Extraction (ITR/E) Project

1.1 Premodifiers of nouns and adjectives (updates to sections 11.1.1 and 11.1.2)

1.1.1 Default constituency within nominals

We assume a default right-branching structure under any NP and NML node. Each daughter of the phrase (whether a single token or itself a constituent node) is assumed to have scope over everything to its right. This means that every daughter also forms a constituent with everything to its right.

This default structure does not apply to NP or NML nodes that have coordinated elements or an apposition structure as daughters, although it can be applied separately to each of those coordinated or appositive elements.

This assumption makes the annotation process for multi-token nominals less complex and the resulting trees more legible, but still allows us to readily derive constituent nodes not explicitly represented. For example, in

```
(NP primary liver cancer)
```

we assume that "liver cancer" is a constituent, and that "primary" has scope over it.

```
(NP a point mutation)
```

"a" has scope over the constituent "point mutation."

So, although we do not show the intermediate nodes in our annotation, our assumed structure for these two NPs would be

```
(NP primary
    (NODE liver
          (NODE cancer)))

(NP a
    (NODE point
          (NODE mutation)))
```

The ability to derive a constituent node for an NP minus its determiner may be useful in aligning syntactic nodes with entities, which may not consistently contain determiners.

1.1.2 The NML node label: marking nominal modifiers

We use the NML node label to mark nominal subconstituents that do not follow the default right-branching structure. Any two or more non-final elements that form a constituent are bound together by NML.

```
(NP (NML Cytochrome P450)
    isoenzyme)

(NP selective
    (NML seratonin reuptake)
    inhibitor)

(NP (NML gel electrophoresis)
    analysis)

(NP (NML 5 nM)
    tetrachlorobiphenyl)

(NP (NML human liver tumor)
    analysis)
```

These nominal subconstituents can contain PPs:

```
(NP the
    (NML (NML guanine)
         (PP to
             (NP cytosine)))
    transformation)

(NP (NML (NML Secretary)
         (PP of
             (NP State)))
    James Baker)
```

Note that NML replaces the use of NAC in nominal modifiers as outlined in section 11.1.2.

1.1.3 Head derivation in NP and NML

The head of an NP or NML is either the rightmost noun (NN or NNS), or is contained within the rightmost NP or NML node.

Recursive applications of this rule can yield the head of phrases containing nested NP or NML nodes. These rules do not apply to coordinated or appositive structures, which have multiple heads, although they can be applied to determine the heads of the individual elements of those structures.

This process of head derivation will not return a head for structures such as (NP the rich) or (NP 12). We can assume either that the NP has a null head, that it's headless, or that JJx or CD can serve as head when a nominal head cannot be derived.

1.1.4 Deverbal adjectives ("New York – based")

NML can also mark multi-token nominal elements modifying an adjective:

```
(NP the
    (ADJP (NML New York)
          -
          based)
    company)

(NP (ADJP (NML Hoe 234)
          induced)
    relaxation)
```

Many circumstances requiring the use of NML have come about due to changes in tokenization policies involving hyphenated items.

```
(NP a
    (ADJP (NML Hong Kong)
          -
          based)
    company)

(NP (NML Taiwan - Palau)
    trade)

(NP (NML (NML Hong Kong)
         -
         (NML Palau))
    trade)
```

1.1.5 Ambiguities within NP

There have been a number of cases where constituency of a complex NP has not been clear to annotators when the information from the entity annotation is unavailable. Here is a representative sampling:

liver cell mutations
p53 gene alterations
ras oncogene expression
polymerase chain reaction
single strand conformation polymorphism

In "p53 gene alterations", for example, it is not clear whether the first two nouns are a constituent ("p53 gene") or the last two are ("gene alterations").

When an annotator comes across one of these cases with multiple alternatives that seem equally viable (and information about the entity annotation of the noun phrase is unavailable), the default is to mark any constituents that are known and leave any ambiguities as flat as possible (following the default right-branching structure):

```
(NP p53 gene alterations)

(NP (NML single strand)
    conformation polymorphism)
```

(if the annotator is certain of the constituency of "single strand" but uncertain of its scope within the phrase).

Treebank constituency follows the domain-specific entity constituency, in the cases where treebank annotators have access to such information.

1.2 Introduction to *P*, placeholder for copied material (addendum to chapters 4 and 8)

Under the Treebank II guidelines, distributed modifiers in coordinated structures have prevented the representation of discontinuous entities as constituents. For example, "K- and N-ras" (representing the entities "K-ras" and "N-ras") could not be tagged to yield constituent nodes associated with "K-ras" and "N-ras".

We have introduced the placeholder for distributed material *P* to solve this problem. *P* is used exclusively in coordinated nominal structures; it is placed in coordinated elements that are missing either a distributed head or a distributed premodifier. In "K- and N-ras", the coordinated premodifier "K-" is missing the distributed head "ras", so the placeholder *P* is inserted after "K-" and coindexed with "ras":

```
(NP (NP K-
       (NML-1 *P*))
    and
    (NP N-
       (NML-1 ras)))
```

This creates constituent nodes "K-ras" and "N-ras" that align with the entities being represented.

The distributed text must be a constituent, and, to allow for coindexation with *P*, must be dominated by an appropriate node. This is true even when the distributed text is single-token, as POS tags are not syntactic nodes and cannot be coindexed. The placeholder *P* is dominated by a matching node, and the two nodes are coindexed.

It is important to note that *P* is not a "trace." *P* does not show syntactic movement; it is a placeholder that shows the distributed reading of some shared material. (As it is a placeholder and not a result of syntactic movement, it is not subject to syntactic restrictions on the environments in which it can occur relative to the material with which it is coindexed.)

Shared material is only copied when its meaning is distributed. If the shared material can be read with each coordinated element with no change in meaning, then it is distributed. If the meaning is changed, it is not distributed, and should be left as a sister to the coordinated structure. (See section 1.3 below on making the collective/distributed distinction.)

Note: Certain uses of *P* replace what would have been *RNR* under the old guidelines. In coordinated NPs, *RNR* is now only used to show shared complements.

1.3 Shared material in coordinated structures (update to parts of chapter 8)

1.3.1 Coordinated nominals sharing premodifiers (update to 8.1.2)

In looking at shared premodifiers in coordinated nominal constructions, we now mark the distinction between distributed and collective premodifiers through the use of copying.

1.3.1.1 Distributed premodifiers

When coordinated nominals share a distributed premodifier, the premodifier is copied so that it forms a constituent with each head.

```
(NP (NP (ADJP-1 enhanced)
        CYP2C9 production)
    and
    (NP (ADJP-1 *P*)
        (NML 11,12 EET)
        production))

(NP (NP (ADJP-1 bronchial)
        washings)
    and
    (NP (ADJP-1 *P*)
        brushings))
```

Multiple distributed premodifiers must each be copied singly to preserve their scope:

```
(NP (NP (ADJP-1 cultured)
        (NP-2 rat)
        pancreatic acinar cells)
    and
    (NP (ADJP-1 *P*)
        (NP-2 *P*)
        hepatocytes))
```

Note the distinction between the above case of multiple single-token premodifiers, which do not form a single constituent and so must be copied separately, and a multi-token premodifier, which does form a constituent and can be copied as a whole:

```
(NP (NP (NML-1 Cytochrome P450)
        inhibitors)
    and
    (NP (NML-1 *P*)
        inducers))
```

"/" is sometimes used as a conjunction, and a premodifier can be distributed across it:

```
(NP (NP (NML-1 CYP)
        4A)
    /
    (NP (NML-1 *P*)
        5A))
```

1.3.1.2 Collective premodifiers

When shared modifiers are collective rather than distributed, they are not copied.

When the coordinated heads are both single-token, the entire structure is left flat:

```
(NP combined washings and brushings)
```

Note that determiners and numerals are considered collective modifiers and are never copied:

```
(NP the dogs and cats)

(NP 11 dogs and cats)
```

Premodifiers in flat coordinated structures are by default collective.

Unshared modifiers must be bound to their head with NP or NML. If one coordinated element is marked with NP or NML, all other elements in the coordinated structure must receive the same node label, even if they have no unshared modifiers (i.e., are single-token):

```
(NP (NP liver cells)
    and
    (NP hepatocytes))
```

In addition, the scope of any premodifiers must be shown when one or more of the coordinated nominals is multi-token (that is, has an unshared or distributed modifier):

```
(NP the
    (NML (NML liver cells)
         and
         (NML hepatocytes)))

(NP the combined
    (NML (NML (ADJP-1 bronchial)
              washings)
         and
         (NML (ADJP-1 *P*)
              brushings)))
```

Note: under the TreeBank II guidelines, scope of prenominal modifiers (both collective and distributive) was shown using NX (8.1.2 #3). The NX node label is no longer used. NML is used to show the scope of collective premodifiers, and distributed modifiers are copied with *P*.

1.3.2 Coordinated modifiers sharing heads on the left

Heads with coordinated post-modifiers (usually numerals) are copied to form a constituent with each post-modifier:

```
(NP (NP (NML-1 codons)
        12)
    and
    (NP (NML-1 *P*)
        13))

(NP (NP (NML-1 exons)
        1)
    and
    (NP (NML-1 *P*)
        2))

(NP (NP (NML-1 CB)
        -52)
    and
    (NP (NML-1 *P*)
        -101))
```

"Codons 11 and 12" provides an illustration of a problem with singular/plural distinctions in distributed constructions. Our treatment yields "codons 12" and "codons 13" while the entities being referred to are clearly "codon 12" and "codon 13." This problem of plural heads distributed to potentially singular constituents is present in all distributed constructions, but it is particularly clear in this pattern, in which each coordinated element is clearly singular.

1.3.3 Coordinated modifiers sharing heads on the right (updates 8.4 and 8.5)

1.3.3.1 Distributed NP/NML head

1.3.3.1.1 Coordinated nominal modifiers

If coordinated nominal premodifiers share a head, we use copying to distribute the head:

```
(NP (NP N-
        (NML-1 *P*))
    and
    (NP K-
        (NML-1 ras)))

(NP (NP c-
        (NML-1 *P*))
    ,
    (NP h-
        (NML-1 *P*))
    ,
    and
    (NP k -
        (NML-1 ras)))

(NP (NP PCR
        (NML-1 *P*))
    and
    (NP (NML Southern Blot)
        (NML-1 analysis)))

(NP the
    (NML (NML ortho
              (NML-1 *P*))
         and
         (NML meta
              (NML-1 positions))))
```

Nominal subconstituents may have distributed heads in addition to the distributed heads of the main NP. These heads should be copied separately to preserve the correct constituent nodes:

```
(NP (NP (NML N-
             (NML-1 *P*))
        (NML-2 *P*))
         and
    (NP (NML K-
             (NML-1 ras))
        (NML-2 genes)))
```

Note that sometimes the distributed text is a multi-token constituent that contains the head:

```
(NP (NP (ADJP naturally occurring)
        (NML-1 *P*))
    and
    (NP (ADJP pharmacologically induced)
        (NML-1 gene mutations)))
```

Also note that coordinated nominal premodifiers are not necessarily distributed, and, if collective, form a constituent marked by NML:

```
(NP the
    (NML quinidine and quinine)
    isomer pair)
```

1.3.3.1.2 Coordinated adjectival modifiers

In deciding how to treat coordinated adjectival modifiers, it is necessary to determine whether they are functioning as intersective or disjunctive modifiers.

Intersective modifiers pick out overlapping sets – that is, sets that have some members in common. They form a constituent node:

```
(NP (ADJP strong and competitive)
    athletes)

(NP (ADJP old and tattered)
    photographs)
```

Disjunctive modifiers pick out non-overlapping sets, and their shared head is distributed to form a constituent node with each:

```
(NP (NP cancerous
        (NML-1 *P*))
    and
    (NP (NML non - cancerous)
        (NML-1 cells)))

(NP (NP old
        (NML-1 *P*))
    and
    (NP new
        (NML-1 photographs)))
```

1.3.3.2 Distributed ADJP head (addendum to Chapter 8)

Coordinated nominal modifiers:

```
(NP (NP (ADJP 11-
              (ADJP-1 *P*))
          (NML-2 *P*))
      and
      (NP (ADJP 12-
              (ADJP-1 hydroxylated))
          (NML-2 proteins)))
```

Coordinated adverbial modifiers:

```
(NP (NP (ADJP very
              (ADJP-1 *P*))
          (NML-2 *P*))
      and
      (NP (ADJP slightly
              (ADJP-1 ripe))
          (NML-2 bananas)))
```

Note these structures typically involve a distributed nominal head, which must be copied separately from the adjective to preserve the correct constituent nodes.

1.3.3.3 Distributed ADVP head

```
(ADVP (ADVP 3
            (ADVP-1 *P*))
        and
        (ADVP 5
            (ADVP-1 fold)))
```

Note: "Fold" has been the only case of a distributed adverb that has actually come up. As we are currently contemplating a change in POS and Treebank guidelines to treat "fold" as a noun, we may not need this category of copying.

1.3.4 Complex cases of distribution

1.3.4.1 Distributed material on the left and right

Phrases with distributed material on both the left and right are fairly common. Multiple uses of copying can produce the correct constituent nodes:

```
(NP (NP (NML (NML-1 cytochrome P-450)
             2D6)
          (NML-2 *P*))
      and
      (NP (NML (NML-1 *P*)
             3A4)
          (NML-2 activity)))
```

```
(NP (NP (NML (ADJP-1 recombinant)
             CYP1A1)
         (NML-2 *P*))
    ,
    (NP (NML (ADJP-1 *P*)
             CYP1A2)
         (NML-2 *P*))
    ,
    and
    (NP (NML (ADJP-1 *P*)
             CYP1B1)
         (NML-2 activities)))

(NP (NP (ADJP-1 standardized)
        NAGE
        (NML-2 *P*))
    and
    (NP (ADJP-1 *P*)
        G115
        (NML-2 extracts)))

(NP (NP (NML (NML-1 CB)
             - 52)
         (NML-2 *P*))
     and
    (NP (NML (NML-1 *P*)
             - 101)
         (NML-2 metabolism)))

(NP (NP (ADJP-1 adrenal)
        (ADJP-2 mitochondrial
        (NML (NML-3 P450)
             (NML 11 beta))
        (NML-4 *P*))
    /
    (NP (ADJP-1 *P*)
        (ADJP-2 *P*)
        (NML (NML-3 *P*)
             18)
        (NML-4 *P*)
    /
    (NP (ADJP-1 *P*)
        (ADJP-2 *P*)
        (NML (NML-3 *P*)
             19)
        -
        (NML-4 hydroxylase)))
```

1.3.4.2 Phrases with coordinated modifiers and coordinated heads

In phrases with coordinated modifiers and coordinated heads, it is impossible to fully distribute all shared elements. For example:

"the N- and K- ras cells and tumors"

contains the entities "N-ras cells", "K-ras cells", "N-ras tumors", and "K-ras tumors". It is impossible to derive all four of these entities with our current mechanism of copying. These cases are relatively rare, and two different annotations can be found, depending upon which level of coordination the distribution is done across:

```
(NP the
        (NML (NML (NML-1 (NML N-
                            (NML-2 *P*))
                         and
                         (NML K-
                            (NML-2 ras)))
                   cells)
              and
              (NML (NML-1 *P*)
                   tumors)))
```

which yields the entities "N-ras and K-ras cells" and "N-ras and K-ras tumors"

and

```
(NP the
        (NML (NML (NML N-
                      (NML-3 *P*))
                  (NML-4 *P*))
             and
             (NML (NML K-
                      (NML-3 ras))
                  (NML-4 cells and tumors))))
```

which yields the entities "N-ras cells and tumors" and "K-ras cells and tumors."

Other phrases that include two levels of coordination:

"N- and K-ras codons 1 and 2"
"androgen and estrogen production and metabolism" [964 #7]

1.3.5 Coordinated constituents sharing adjuncts on the right (update to 8.3.2)

Whenever possible, shared adjuncts are adjoined to coordinated structures:

"5 and 10 mg/ L"

```
(NP (NP (NP 5
            (NML-1 *P*))
        and
        (NP 10
            (NML-1 mg)))
    (PP /
        (NP L)))
```

Note that a similar policy is used for shared adjuncts in VP, except that the adjuncts are left at coordination level rather than adjoined to it. (see 8.3.1)

In cases where coordinated nominal premodifiers in NP share a preposition, the elements the preposition modifies do not end up as sisters to each other. In these cases, the PP should be copied so that it adjoins each coordinated element:

"5 and 20 mg/kg b.w. melatonin" [821#9]

```
(NP (NP (NML (NML 5
                  (NML-1 *P*))
             (PP-2 *P*))
        (NML-3 *P*))
    and
    (NP (NML (NML 20
                  (NML-1 mg))
             (PP-2 /
                   (NP kg b.w.)))
        (NML-3 melatonin)))
```

1.3.6 Restrictions on Copying

It is important to note that copying is only used to show the distribution of shared material in coordinated nominals. To avoid overapplication of copying to all cases of entity/constituent misalignment, it is important to note when copying is not used.

1.3.6.1 Gapping (update to 7.4.1) and Coordinated VP: use *RNR*

Copying should never be used in gapped structures or regular coordinated VPs. Use *RNR* instead:

```
(VP were
    (VP (VP found
            (NP-2 *)
            (PP-LOC=1 in
                      (NP normal
                          (NML-3 *RNR*))))
        but
        (VP not
            (PP-LOC=1 in
                      (NP cancerous
                          (NML-3 *RNR*))))
        (NML-3 cells)))

(VP (VP (ADVP weakly)
        inhibited
        (NP basal
            (NML-1 *RNR*)))
    and
    (VP (ADVP strongly)
        inhibited
        (NP basolateral
            (NML-1 *RNR*)))
    (NML-1 cells))
```

1.3.6.2 PP coordination: use *RNR*

We do not copy within coordinated PPs. Use *RNR* to show the distribution of all shared material.

```
(VP ranged
    (PP (PP from
            (NP (NP 5
                    (NML-1 *RNR*))
                (PP-2 *RNR*)))
        (PP to
            (NP (NP 6
                    (NML-1 *RNR*))
                (PP-2 *RNR*)))
        (NML-1 mol)
        (PP-2 /
              (NP L))))
```

```
(PP-LOC (PP in
            (NP renal
                (NML-1 *RNR*)))
        and
        (PP in
            (NP hepatic
                (NML-1 *RNR*)))
        (NML-1 cells))
```

1.3.6.3 Do not copy from lower to higher position

Copying is not used to distribute material from a PP to an NP that adjoins it. Leave the NP headless.

```
(NP (NP parallel)
    (SBAR as
          (S (NP-SBJ-1 *)
             (VP opposed
                 (NP-1 *)
                 (PP-CLR to
                         (NP proportional inhibition)))))
```

"higher central than peripheral oestrogenic activity" [963#2]

```
(NP (NP higher central)
    (PP than
        (NP peripheral oestrogenic activity)))
```

A related case is the "4 of 5 cats" or "4/5 cats" pattern, in which copying or *RNR* are never used:

```
(NP (NP 4)
    (PP of
        (NP 5 cats)))
```

Note also that some patterns of this type can be annotated as coordination rather than adjunction. For example, "versus" can be treated as a conjunction rather than a preposition, in which case copying can be used:

```
(NP (NP (NML childhood
             -
             (NML-1 *P*))
        (NML-2 *P*))
    versus
    (NP (NML adult
             -
             (NML-1 onset))
        (NML-2 disease)))
```

```
(NP (NP (NP (NP 85
                (NML-1 *P*))
            versus
            (NP 17
                (NML-1 %)))
        (PP-LOC in
            (NP the adenomas)))
    , and
    (NP (NP (NP 50
                (NML-2 *P*))
            versus
            (NP 10
                (NML-2 %)))
        (PP-LOC in
                (NP the follicular carcinomas))))
```

If the material shared across different levels is itself a PP, the use of *RNR* to show the distribution is sometimes seen:

```
(NP (NP (NP 10 micrograms)
        (PP-1 *RNR*))
    (PP –
        (NP (NP 10 mg)
            (PP-1 *RNR*)))
    (PP-1 /
          (NP ml)))
```

1.3.6.4 Exception: Copying allowed for the "Codons 2-12" pattern

We do allow copying across PP when there is a numeral or numeral-like range following the head, as in the phrase "codons 2-12"

```
(NP (NP (NML-1 codons)
        4)
    (PP –
        (NP (NML-1 *P*)
            12)))

(NP (NP (NML-1 stages)
        4a)
    (PP –
        (NP (NML-1 *P*)
            5b)))
```

1.3.6.5 Coordinated adjuncts

We do not show the distribution of heads among coordinated adjuncts.

```
(NP (NP simultaneous mutation)
    (PP of
        (NP (NP ras genes)
            and
            (NP the p53 gene))))
```

1.3.7 Coordinated constituents sharing complements on the right (reiteration of 8.2)

TreeBank II policy on shared complements (see 8.2) is unchanged. Complements are left at coordination level when the coordinated elements are single token and *RNR*-attached inside each element when one or more elements are multi-token. This applies to shared complements of nouns, adjectives, verbs, and prepositions.

1.4 Recursivity of relative clause adjuncts to NP (update to 11.2.3)

Relative clauses, both reduced and full, restrictive and non-restrictive, are now recursively adjoined to an NP that includes everything that comes before them:

i.

```
(NP (NP (NP the President)
        (PP of
            (NP the United States)))
    (SBAR (WHNP-1 who)
          (S .... )))

(NP (NP (NP the book)
        (PP about
            (NP toads)))
    (SBAR that is on the table)
```

Note that PPs following a relative clause are not recursively adjoined:

ii.

```
(NP (NP the book)
    (SBAR that is on the table)
    (PP about toads))
```

Note that there is an asymmetry between examples (i) and (ii). We are tolerating this for the time being, but annotators have been instructed to make a note in the comments field when they encounter pattern (ii). The alternative structure that may be used in the future is

```
(NP (NP (NP the book)
        (PP-1 *ICH*)
    (SBAR that is on the table)
    (PP-1 about toads))
```

1.5 Parentheses containing acronyms and renamings (addendum to 2.6)

When the material enclosed in parentheses can be interpreted as an appositive, we do not use the PRN node but instead adjoin the entire parenthetical to the preceding nominal. Acronyms and renamings are treated in this manner:

```
(NP (NP warfarin)
    and
    (NP (NP acenocoumarol)
        (NP (4'-nitrowarfarin) )))

(NP (NP polymerase chain reaction)
    (NP (PCR) ))

(NP the (NML (NML polymerase chain reaction)
             (NML (PCR) ))
        analysis)
```

See "Section 7: Parentheticals" for information on the annotation of other types of parentheticals.

1.6 NP, the “namely” problem

Namely/possibly/notably/especially/in particular/e.g. These all go as ADVP/PP inside the NP they introduce. They do not add extra structure.

```
(NP (NP some second-generation antihistamines)
    (NP (ADVP notably)
        terfenadine and astemizole))
```

Sometimes, the phrase introduced by the ADVP needs to be *ICH*ed to adjoin to the correct phrase:

```
(S (NP-SBJ-2 (NP clonal chromosomal abnormalities)
             (NP-1 *ICH*))
   (VP were
       (VP noted
           (NP-2 *)
           (PP in
               (NP (NP 12)
                   (PP of
                       (NP 19 patients))))
           ,
           (NP-1 (ADVP most commonly)
                 - 7))))  [493.9]
```

Unresolved issue:

Some constructions do not lend themselves to this appositive treatment:

"to maintain vascular tone in humans under conditions of a failing nitric oxide synthesis, e.g. in atherosclerosis" [633.8]

1.7 Anti-NN

"anti + NN" constituents are tagged as NML rather than ADJP.

```
(NP (NML anti
         -
         CNTF)
    antibodies)
```

If "anti" needs to be copied, it should be tagged ADVP.

```
(NP (NP (NML (ADVP-1 anti)
             -
             CNTF)
         (NML-2 *P*))
    and
    (NP (NML (ADVP-1 *P*)
             -
             SMN)
         (NML-2 antibodies)))
```

1.8 Chemical names

In general, leave chemical names flat, except when there are clear adjectival or prepositional components to the name. For example:

```
(NP (ADJP Pentane 1,3,5-tricarboxylic)
    acid)
```

1.9 "/" and single- versus multi-token NPs

"/" is sometimes interpreted as a conjunction, and the normal rules of coordination apply. No internal structure is needed if all the coordinated elements are single-token, but each element must be marked if one is multi-token:

"Multimodal cancer treatment mediated by a replicating oncolytic virus that delivers the oxazaphosphorine/rat cytochrome P450 2B1 and ganciclovir/herpes simplex virus thymidine kinase gene therapies." [658 #2]

```
(NP (NML (NML ganciclovir)
         /
         (NML (NML herpes simplex)
              virus))
    (NML thymidine kinase)
    gene therapies)
```

Note that in this particular example, copying is not required, but other uses of "/" may require it. For example "CYP4A/5A" refers to distinct enzymes, and "CYP" must be copied to make "CYP5A" a constituent:

```
(NP (NP (NML-1 CYP)
        4A)
    /
    (NP (NML-1 *P*)
        5A))
```

2 Empty Categories

2.1 *EXP* Passives with It-Extraposition

In our treatment of the construction we informally call "pseudo-passives," we insert both it-extraposition and a passive trace:

```
(S (NP-SBJ-1 (NP It)
             (SBAR-2 *EXP*))
   (VP is
       (VP reported
           (NP-1 *)
           (SBAR-2 that the moon orbits the earth))))
```

We also apply this treatment to non passive versions of the above, e.g. "I like it that X."

```
(S (NP-SBJ I)
   (VP like
       (NP (NP it)
           (SBAR-1 *EXP*))
       (SBAR-1 that X)))
```

2.2 Coindexation of Adverbial Null Subjects

Null subjects are not coindexed across an NP, SBAR, or PP barrier. This explains why we don't coindex reduced or wh- relative clauses, for example. In keeping with previous policy, we still coindex the null subject of S-ADV at main VP or S level when appropriate:

```
(S (NP-SBJ-1 I)
   (VP threw
       (NP the knife)
       (S-ADV (NP-SBJ-1 *)
              (VP killing
                  (NP the king)))))

(S (NP-SBJ-1 We)
   (VP lowered
       (NP the threshold)
       (S-ADV (NP-SBJ-1 *)
              (VP to
                  (VP remove
                      (NP X))))))
```

In the following sentence, the subject position of the infinitival is NOT coindexed:

K-ras was used to analyze the mutation.

```
(S (NP-SBJ-1 K-ras)
   (VP was
       (VP used
           (NP-1 *)
           (S-PRP (NP-SBJ *)
                  (VP to analyze the mutation)))))
```

This is because the coindexed reading "K-ras analyzes the mutation" does not work. Note that in this case the infinitival is adverbial. It is not a complement. So we can still insert a separate passive trace without having to consider "use" a ditransitive verb.

3 Gapping

3.1 Format of Gap Indices

A change in the format of gapping indices has been made. In the original treebank, the elements in the template were marked with "-" hyphen indices, while the elements in the following gapping conjuncts were marked with "=" equals indices. However, this indexing format interacts badly with any non-gapping indexing that may be necessary in a tree (passive traces, for example, or extraposition or right node raising). For that reason, it is now the case that *all* gapping elements are marked with "=" indices. So, in the following example, the NPs for *both* Bach (in the template) and Beethoven (in the second conjunct) are now marked "=2":

```
(S (S (NP-SBJ=1 Mary)
      (VP likes
          (NP=2 Bach)))
   and
   (S (NP-SBJ=1 Susan)
      ,
      (NP=2 Beethoven)))
```

3.2 Use of Gapping with "but not"

When the negation in a conjunction such as "but not" can possibly be assigned to a verb, annotate as gapped VP:

```
(S (NP-SBJ-1 X)
   (VP was
       (VP (VP found
               (NP-1 *)
               (PP-LOC=2 in
                       (NP Y)))
           but
           (VP not
               (PP-LOC=2 in
                         (NP Z))))))
```

3.3 New policy on the use of the anti-placeholder *NOT*

The *NOT* anti-placeholder is used only in the template (never in the copy) to mark a node that is not copied from the template. This reflects a strict interpretation of template-to-copy copying: everything in the template is interpreted in the copy, with any gap-coindexed piece substituted by the corresponding piece in the copy. *NOT* is required only when there is something in the template that is not interpreted in the copy. If the copy has an element that is not in the template, no additional annotation is required.

A corollary to this is that we do not have to put the anti-placeholder in for the passive trace in the "was found in Y but not in Z" examples. The passive trace is assumed to be interpreted in the copy, even though it has not been explicitly tagged that way.

Here's an example from the corpus, Sanger 36 #11.

Note that we don't put *NOT* in the template and copy #2 for "namely" based on these rules. Also, we are assuming that "for the first time" is interpreted in both copies in addition to the template (if it weren't, we would need *NOT*).

```
(S (PP In
       (NP addition))
   ,
   (NP-SBJ-1 the
             (NML codon 560)
              mutation)
   (VP could
       (VP be
           (VP (VP demonstrated
                   (NP-1 *)
                   (PP-TMP for
                           (NP the first time))
                   (PP=2 in
                         (NP indolent mastocytosis)))
               ,
               (VP (ADVP namely)
                   (PP-LOC=2 in
                             (NP (NP two)
                                 (PP of
                                     (NP (NP four specimens)
                                         (PP from
                                             (NP adult
                                                 patients)))))))
               ,
               but
               (VP not
                   (PP-LOC=2 in
                             (NP (NP those)
                                 (PP from
                                     (NP two children)))))))
   .)
```

Another example, file 661 #13. Note also the use of the ADVP tag on "not", usually left untagged, so that it can receive a gap coreference to provide the correct positive reading for "cytotoxic only at high concentrations."

```
(S (NP-SBJ It)
   (VP (VP was
           (ADVP=5 not)
           (NP-ADV itself)
           (ADJP-PRD=1 genotoxic)
           (PP-LOC=2 in
                     (NP hepatocytes))
           ,
           (SBAR-ADV=3 as
                       (S (NP-SBJ-4 *)
                          (VP measured
                              (NP-4 *)
                              (PP-MNR by
                                      (NP unscheduled DNA
                                          synthesis))))))
       ,
       or
       (VP (ADJP-PRD=1 mutagenic)
           (PP-LOC=2 in
                     (NP (NP (NP the strain)
                             (PP of
                                 (NP Salmonella)))
                         (VP employed
                             (NP *))))
           (SBAR-ADV=3 *NOT*))
       and
       (VP (ADVP=5 *NOT*)
           (ADJP-PRD=1 cytotoxic)
           (PP (ADVP only)
               at
               (NP high concentrations
                   (PRN ((NP (QP > or = 0.5)
                             mM)))))
           (PP-LOC=2 *NOT*)
           (SBAR-ADV=3 *NOT*)))
   .)
```

3.4 Level of Gap Coindices

Elements receiving a gap coindex do not have to be daughters of the gapped phrase, that is, VP, S, or SBAR.

```
(S (NP-SBJ (NP alteration)
           (PP of
               (NP (NP a substrate binding site)
                   (PP of
                       (NP P-450MP)))))
   (VP may
       (VP (VP reduce
               (NP its ability
                   (S (NP-SBJ *)
                      (VP to
                          (VP hydroxylate
                              (NP=1 S-mephenytoin))))))
           but
           (VP not
               (NP=1 tolbutamide)))))  [644, #10]

(VP (VP oxidized
        (NP *)
        (PP-CLR to
                (NP X))
        (PP by
            (NP (NP washed microsomes)
                (PP from
                    (NP=1 (NP (NP (NML-1 rat) liver)
                              and
                              (NP (NML-1 *P*) kidney)
                          ,
                          and
                          (NP rabbit))))))
    but
    (VP not (NP=1 rat lung))) [1017 #5]
```

Note, however, that a gap coindex cannot be nested inside another gap coindex.

```
(S (S (NP-SBJ (NP the apparent km)
              (PP=1 for
                 (NP (NP the aromatization)
                     (PP of
                         (NP andostenedione)))))
      (VP is
          (NP-PRD=3 95nM)))
   and
   (S (PP=1 for
            (NP 16alpha-hydroxytestosterone))
      (NP-PRD=3 7muM)))
```

An annotator might be tempted to put an additional gap coindex on "andostenedione" and "16alpha-hydroxytestosterone" to yield "the apparent km for the aromatization of 16alpha-hydroxytestosterone" as the subject of the copy. However, we do not use such additional gap coindices, as they lead to confusion as to exactly which material is interpreted in the copy.

3.5 Gapping at SBAR level

Gapping can be done at SBAR level:

```
(NP (NP seven mutations)
       (SBAR (SBAR (WHNP-1 (NP=2 five)
                           (WHPP of
                                 (WHNP which)))
                   (S (NP-SBJ-1 *T*)
                      (VP were
                          (NP-PRD=3 frameshift mutations))))
                    and
             (SBAR (NP=2 two)
                   (S (VP were
                          (NP-PRD=3 nonsense ones))))))
```

Note the gap-coindices on "two" and "five" to preserve the correct reading of "two of which were nonsense mutations" in the copy.

3.6 Gapping occurs at the level of the main verb rather than at auxiliary level

We do not deal with the level of modification of adverbs in gapped structures. That is, we do not make the distinction between auxiliary-level modification and modification at the level of the main verb. Given the option, we don't include auxiliaries in gapped structures, and gap at the level of the main verb whenever possible.

```
(VP was
    (VP (VP found
            (NP-2 *)
            (PP=1 in X))
        but
        (VP not
            (PP=1 in Y))))

(VP was
    (VP (VP stimulated
            (NP-1 *)
            (ADVP-EXT=2 (QP 1.5 fold))
            (PP=3 in X))
        and
        (VP (ADVP-EXT=2 (QP 2.5 fold))
            (PP=3 in Y)))
```

3.7 Gapping and shared material, use *RNR*

The placeholder *P* is never used across coordinated or gapped VP, S, or SBAR. We use *RNR* to show the distribution of shared material.

RNR can be used for elided NP heads:

```
(VP (VP were
       (VP found
           (NP-1 *)
           (PP-LOC=2 in
                     (NP normal
                         (NML-3 *RNR*)))))
    but
    (VP not
        (PP-LOC=2 in
                  (NP cancerous
                      (NML-3 *RNR*))))
    (NML-3 cells))
```

PP adjuncts can also be RNRed if the annotator feels strongly that they should be interpreted in multiple positions, as in the following example from file 823, #23:

```
(VP (VP decreased
        (NP X)
        (PP by
            (NP (NP factors)
                (PP of
                    (NP=2 0.17))))
        (PP-TMP=3 on
                  (NP (NP the first day)
                      (PP-1 *RNR*))))
    and
    (VP (NP=2 0.14)
        (PP-TMP after
                (NP (NP one week)
                    (PP-1 *RNR*))))
    (PP-1 of
          (NP treatment))
    (ADVP respectively))
```

4 Function tags

Update on use of function tags:

4.1 -TMP

-TMP is used when a phrase specifies a point in time, or a duration. Because -TMP is an adverbial marker, it technically has to be in a position where the verb can "see" it, i.e. daughter of VP. In reality, we thwart this rule all the time, as it is sometimes unavoidable, as with -LOC inside of NP. What this does mean is that we put the -TMP tag as close to the verb as we can. The -TMP tag is not put on a complement of PP, but rather on the PP itself:

```
(PP-TMP over
        (NP five years))
```

4.1.1 Nested -TMP

It is permissible to put in nested -TMP tags when different events are being referenced:

```
(PP-TMP at
        (NP (NP various times)
            (PP-TMP after
                    (NP (NP the removal)
                        (PP of
                            (NP the implant))))))
```

4.2 -MNR

We have been using the -MNR to refer to both manner ("impatiently") and instrument ("by PCR", "using PCR").

4.3 -LOC

Physical things (skin, tumors, patients) always take -LOC, in both abstract and specific usages. For example, "in human skin" takes -LOC, even though that refers generically to human skin rather than to a concrete object. We are including genes, codons, and DNA in the category of physical things.

Diseases in the abstract do not take -LOC. "In melanoma", "In lung carcinoma". However, many disease names can easily be made concrete through the addition of a determiner or a plural marker, in which case they do take the -LOC tag. "In melanomas", "in the lung carcinoma observed."

4.4 -NOM

-NOM is only used with free relatives and gerunds. It is not used with SBAR "whether" or "that", even when this SBAR functions as the subject of the sentence.

```
(SBAR-SBJ Whether
          (S ... ))
```

4.5 -CLR

The -CLR tag is used mainly to indicate prepositional phrases that can be construed as arguments for a particular sense of a verb. The following list represents all uses of the -CLR tag in the first 300 files of the oncology corpus. Those considered “questionable” by the annotators are marked (?) and are likely to be less consistently marked -CLR in the corpus. This list should by no means be considered exhaustive:

account for
act as
act on
aim at (?)
arise from
base on/upon
bestow to/on
bind to
associated with
classify as
code for
combine with
compare to
compare with
composed of
confine to
consist of
contribute to
convert to
correlate with
correspond to
couple with
define as
depend on
describe as (?)
derive from
died of (?)
distinguish from
encode for
expose to

extract from
function as
give rise to
known as
hybridize to/with (?)
implicate in (?)
interact with
interfere with
involve in
lead to
limit to
link to
look at
look for
made up of
maps to
participate in (?)
point to
predispose to
present as (The disease presents itself as X)
progress to
react to/with
restrict to
refrain from
related to
report on
result in/from
screen for (?)
search for
serve as (?)
suffer from
take part in

Note that occasionally the -CLR tag is put on an NP in certain phrasal verbs such as "take part in."

```
(VP take
    (NP-CLR part)
    (PP-CLR in
            (NP the race)))
```

We have also extended the use of -CLR on the ADVP tag to include the "ranked fifth" pattern:

```
(S (NP-SBJ He)
     (VP ranked
         (ADVP-CLR fifth)
         (PP in
                (NP the finals))))
```

4.6 -PUT

The locative complement of the verb "place" takes the -PUT tag:

```
(VP place
    (NP the book)
    (PP-PUT on
            (NP the table)))
```

4.7 Multiple function tags

Adverbial and predicate tags can be combined as long as the -PRD tag is last (S-PRP-PRD):

```
(S (NP-SBJ It)
   (VP was
       (PP-LOC-PRD in
                   (NP the store))))

(S (NP-SBJ (NP The aim)
           (PP of
               (NP this study)))
   (VP was
       (S-PRP-PRD (NP-SBJ *)
                  (VP to
                      XYZ ))))
```

When combined with another function tag, the -SBJ tag must also come last. For example, use S-NOM-SBJ rather than S-SBJ-NOM.

5 Verbs: Pseudo-Prepositions, making the Adjective-Verb distinction

"Combined with" and "compared with" have been dropped from the list of pseudo-prepositions.

In general, treat items as verbs whenever possible (instead of as pseudo-prepositions in the above two cases or as adjectives in many others).

```
(S (NP-SBJ-1 These results)
   (VP were
       (VP associated
           (NP-1 *)
           (PP-CLR with
                   (NP X)))))
```

6 NAC

NAC is no longer used to indicate pre-head modifiers, but is still used around conjunction plus adjunct patterns when ordinary coordination is impossible. These patterns are basically adverbials, but cannot be tagged as such because of the extraneous conjunction. A good test for NAC is to see if the sentence reads well without the conjunction.

For example, "We ran, but slowly" works as "We ran, slowly" and "X was associated with Y, and more so in Z than in ZZ" (877 #10) works as "X was associated with Y, more so in Z than in ZZ". Both of these can be tagged as NAC:

```
(S (NP-SBJ-1 X)
   (VP was
       (VP associated
           (NP-1 *)
           (PP-CLR with
                   (NP Y)
           ,
           (NAC and
                (ADVP (ADVP more so
                            (PP-LOC in
                                    (NP Z)))
                      (PP than
                          (PP-LOC in
                                  (NP W))))))))
```

"X activity was also reduced by fluid from dominant follicles, but not to a greater extent than in basal conditions." [899 #9]

```
(VP was
    (ADVP also)
    (VP reduced
        (NP-1 *)
        (PP by
            (NP-LGS (NP fluid)
                    (PP from
                        (NP dominant follicles))))
        ,
        (NAC but
             not
             (PP to
                 (NP (NP a greater extent)
                     (PP than
                         (PP in
                             (NP basal conditions)))))))
```

Earlier examples in the manual for the use of NAC with "but slowly" and "but not very much."

7 Parentheticals

Note that with our standoff annotation tools the annotators now see the punctuation as it is in the source, and we are therefore showing the actual parentheses and other punctuation marks in the examples in this addendum. However, in our released merged files, the punctuation brackets from the source text are replaced by -LRB- (for "(", or "Left Round Bracket"), etc., as in the original Penn Treebank files.

7.1 FRAG and Parentheticals

In this corpus, there are often various random things thrown together in a parenthetical. The relationship between these pieces is represented whenever possible. For example, in the parenthetical "(mean age, 67)", the two NPs are tied together as an S, as there is obvious subject-predicate relationship between them:

```
(PRN ((S (NP-SBJ mean age)
         ,
         (NP-PRD 67))))
```

FRAG basically means that there is no argument structure connecting the daughter phrases together, but that there a relationship that cannot otherwise be defined syntactically. If there are multiple elements within a PRN, they do not necessarily need to be held together with FRAG. FRAG is used if the parts are not completely disparate, and there is some logical relationship binding them together:

```
(PRN ((FRAG (NP San Francisco)
            ,
            (NP California))))

(PRN ((FRAG (S (NP-SBJ p)
               (VP =
                   (NP .05)))
            ,
            (NP chi square test))))
```

FRAG has a different meaning when it is used to mark off material that we do not treebank, such as the by-line, citation, and PMID info at the beginning and end of each abstract. This is a fundamentally different use of FRAG than those discussed above, but can be clearly disambiguated from context: this second usage of FRAG is found in sections rather than sentences, and contains no other syntactic nodes inside it. In its standard use, FRAG is not used to bind together arbitrary material; there has to be some relationship between the elements being bound.

7.2 Citations within the text

Citations within the main text of an article should be marked as FRAG with no internal structure. A PRN node should be used when the citation is enclosed in parentheses.

```
(PRN ((FRAG Shelton et al., 1983)))
```

7.3 The PRN/Acronym Distinction

7.3.1 TYPE 1. Acronyms, and other renamings

Treat as appositives. Do not use PRN node.

Citation of abbreviations in parentheses comes up a lot in this corpus, and seems to be fundamentally different from other types of parentheticals, so much so that there is a strong case for dropping the use of the PRN node label, which is used to indicate material that is syntactically removed from the main sentence. Having the acronym form a constituent with the original term is very compelling given their close relationship.

```
(NP (NP hepatocellular carcinoma)
    (NP (HCC) ))
```

When the acronym or renaming covers only part of the overall NP, NML is used:

```
(NP 34
    (NML (NML hepatocellular carcinoma)
         (NML (HCC))
    specimens)

(NP benign
    (NML (NML bile duct proliferations)
         (NML (BDPs) )))

(NP Oesophageal
    (NML (NML gastrointestinal stromal tumors)
         (NML (GISTs) )))

(NP the quail
    (NML (NML preoptic area)
         (NML (POA) )))
```

An appositive following an acronym is recursively adjoined:

```
(NP (NP (NP endometrial cancer)
        (NP (EC) ))
    ,
    (NP a
        (ADJP very common)
        female pelvic malignant disease))
```

The parentheses can also contain the longer version of the name, and the main text the acronym:

```
(NP (NP PCR)
    (NP (polymerase chain reaction) ))
```

7.3.2 TYPE 2. Numbers, percentages, explanatory notes

For all other parentheticals, a PRN node is used. Because of the flexibility in the placement of these parentheticals, it is difficult to determine a consistent policy for parenthetical adjunction, so the PRN node is never adjoined to anything – no extra structure is inserted into the main sentence. As much structure as possible is put in inside the PRN node, and it is attached at the level that makes the most semantic sense.

```
(NP-SBJ (NP 15
            (PRN (
                 (NP 83%)
                  ) ))
        (PP of
            (NP the 18 SPNs)))

(PP-LOC in
        (NP codon 32
            (PRN ((NP 5 cases)))))
```

Examples showing the variation in parenthetical placement:

4 (80%) of 5 patients
4 of 5 (80%) patients
4 of 5 patients (80%)

80% (4/5) of patients
80% of patients (4/5)

Enumeration also falls under this type:

several alkylphenols (e.g., 2,6-di-tert-butyl-4-methylphenol, BHT)

Note that what falls inside the PRN is syntactically separate from the rest of the sentence. Thus, nominal contents are always tagged NP, even if the PRN falls inside of a NML.

```
(NP (NP 4)
    (PP of
        (NP 5
            (PRN (NP 80%) )
            patients)))
```

Note also that when a parenthetical clearly is associated with a single-token element in the main sentence, extra structure may need to be inserted so it can be attached at the correct level:

```
(NP (NP (ADJP proximal
              (PRN ((NP (NP gastroesophageal junction)
                        or
                        (NP cardia)))))
        (NML-1 *P*))
    and
    (NP (ADJP distal
              (PRN ((NP body or antrum))))
        (NML-1 gastric tumors)))
```

7.3.3 Interaction between parentheticals of Types 1 and 2

"36 medullary thyroid carcinomas (MTCs) (16 hereditary and 20 sporadic) and ten pheochromocytomas (eight hereditary and two sporadic)" [85#5]

```
(NP 36
    (NML (NML medullary thyroid carcinomas)
         (NML MTCs))
    (PRN (NP (NP 16 hereditary)
             and
             (NP 20 sporadic)))
```

Parentheticals of both types can be enclosed in one set of parentheses, but they are split up in annotation:

"phenobarbital (PB; 0.05%)" [1020#5]

```
(NP (NP phenobarbital)
    (NP ( PB )
    (PRN ; (NP 0.05%)) ))
```

The parenthetical is serving two functions here. The first part, "PB", provides an acronym for "phenobarbital" and under our current system is treated as an appositive. "0.05%" is ordinary parenthetical material, tagged PRN.

7.4 Sample annotation of parenthetical data citations

To aid in consistent tagging of parentheticals, here are a few sample data citations whose treatment we've agreed on:

```
(PRN ( (FRAG (NP (NP 13)
                 (PP /
                     (NP29)))
             ,
             (NP 45%)) ) )

(PRN ( (FRAG (NP 8 subjects)
             (S (NP-SBJ age range)
                (NP-PRD (QP 42-55)))) ) )

(PRN ( (S (NP-SBJ age range)
          (NP-PRD (QP 23-71)
                  years)) ) )

(PRN ( (S (NP-SBJ (NML 90 %)
                  CI)
          (NP-PRD (QP 0.05 - 0.36))) ) )
```

In general, we are treating terms that have an implied copular relationship as subject and predicate inside an S. Terms that are related in other ways ("13/29, 45%", or "8 subjects, age range 42-55") are bound together by FRAG.

7.5 Parentheticals not introduced by parentheses

In general, material that is not set off by parentheses should NOT be tagged as PRN. There may very occasionally be asides not offset by parens that are so tangential that they cannot conceivably fit into the syntax of the overall sentence. These can take the PRN tag, but will probably not come up very often in the ITR/E corpus.

For example, "is different, at least in proportion and/or in nature, from X", does not take PRN, as the "different in X from Y" pattern is perfectly acceptable.

(Note again that not all material in parentheses is tagged PRN, as in the case of appositives and acronyms.)

8 Numbers

8.1 QP (update to 11.3.1)

Words such as "to" and symbols such as "-" and "x" are left flat when they are part of complex quantifier phrases:

```
(NP (QP 12 to 13)
    %)

(NP (QP 63 to 69)
    %)

(NP (NML (QP 5 - 6)
         mg)
    morphine)

(NP (QP 5 x 10(-7))
    mols)
```

When a complex quantifier contains one or more multi-token quantities, every quantity in the phrase must be marked with a nested QP node. The words and symbols expressing the relationship between the quantities are still left flat:

```
(NP (QP (QP 10(-7))
        -
        (QP 4 x 10(-6)))
    mol)
```

8.2 Scientific notation

Scientific notation takes a QP label:

```
(NP (QP 1.2 x 10(6))
    cells)
```

8.3 Nested QP

Multi-token quantities inside a larger QP get marked as QP, following standard coordination-marking rules:

```
(NP (QP (QP 5 x 10(-7))
        -
        (QP 4 x 10(-6)))
    mol)
```

```
(NP (QP (QP 10(-7))
        -
        (QP 4 x 10(-6)))
    mol)
```

Note that "-", "/", and other mathematical operators never get structure inside of QP. (They are sometimes treated as prepositions in other environments.)

In times/fold constructions, annotators will sometimes put in nested QPs:

```
(NP a
    (QP (QP 2 - 3)
        fold)
    increase)
```

8.4 Discontinuous QP

The "2.5 mg or greater" pattern is annotated as a discontinuous QP:

```
(NP (QP 2.5)
    mg
    (QP or greater))
```

If the second portion of the QP falls at a different level of the tree, it can be *ICH*-attached to the correct level:

```
(NP (NP (QP 2.5
            (QP-1 *ICH*))
        micrograms)
    (PP /
        (NP ml))
    (QP-1 or greater))
```

8.5 Complex Fractions

Complex fractions are often not nested. For example:

```
(NP (NP .018 mumol)
    (PP /
        (NP hr))
    (PP /
        (NP 10(6) cells)))
```

8.6 “Times” and “fold”

"Times" and "fold" are both put inside of QP:

```
(NP (QP three times)
    the amount)

(NP a
    (QP three fold)
    increase)
```

When they are not in prenominal position, the node label dominating QP follows from the POS tags for times/NNS and fold/RB:

```
(VP rose
    (NP-EXT (QP 3 times))

(VP increased
    (ADVP-EXT (QP 3 fold)))

(VP increased
    (PP-EXT by
            (ADVP (QP 3 fold))))

(VP increased
    (ADVP-EXT (QP (QP 1.5
                      (ADVP-1 *P*))
                  and
                  (QP 3.5
                      (ADVP-1 fold))))
    (PP-LOC in
            (NP X and Y))
    (ADVP respectively))
```

9 Ranges and Endpoints Summary

In prenominal position ranges are annotated as QP:

```
(NP (QP from 10 to 20)
    mg)

(NP (QP 10 - 20)
    nm)

(NP (QP 10 to 20)
    %)

(NP a
    (NML (QP 10 - 20)
         %)
    increase)
```

Note that we have no special treatment for dimensionless quantitative values (%, fold, etc.).

Also note that endpoints in prenominal position do get the full PP structure:

```
(NP the
    (NML (NML G)
         (PP to
             (NP C))
    transversion)
```

When "from... to..." ranges or endpoints post-modify a verb or noun, they are annotated as PP. Ranges form a constituent, and endpoints do not. In both patterns, we use *RNR* for distributed material:

```
(VP increased
    (PP from
        (NP 10
            (NML-1 *RNR*)))
    (PP to
        (NP 20
            (NML-1 *RNR*)))
    (NML-1 %)
```

```
(VP ranged
    (PP (PP from
            (NP 10
                (NML-1 *RNR*)))
        (PP to
            (NP 20
                (NML-1 *RNR*)))
        (NML-1 %)))
```

An actual unit like "mg" gets the same treatment as the dimensionless "%".

10 Adjectives

10.1 Determining Adjectival complements

A PP that follows a predicative adjective is either a complement to that adjective or attached at some other level (typically, the copula). That determination can be difficult to make.

If the adjective is deverbal and the PP contains an NP that could be considered an argument of the verb the adjective is derived from, then consider the PP to be a complement to the adjective. If not, then it might not be a complement (but note that non-deverbal adjectives can take complements sometimes too).

For example, in the following two cases, the PP headed by "of" is a complement (as "of" almost always is), but the second the PP headed by "among" is attached to the copula.

```
(S (NP-SBJ (NP The sole presence)
           (PP of
               (NP K-ras mutation)))
   (VP was
       (ADJP-PRD predictive
                 (PP of
                     (NP survival)))))

(S (NP-SBJ The effect)
   (VP may
       (VP be
           (ADJP-PRD different)
           (PP among
               (NP various countries)))))
```

More generally, we take a strict approach to what can be attached to adjective level, attaching only phrases that are non-optional for the given sense of the adjective.

10.2 JJ to JJ, RB to RB

```
(NP the
    (ADJP apical
          (PP to
              (ADJP basolateral)))
    flux)
```

```
(NP a
    (ADJP (ADVP mucosally
                (PP to
                    (ADVP serosally)))
          directed)
    K+ gradient)
```

11 Miscellany

11.1 Conjunction of similar elements

When S and SINV are conjoined, they are bound by S rather than UCP:

```
(S (S X does not verb1)
   nor
   (SINV does it verb2))
```

This also applies similarly to NP and S-NOM:

```
(NP (S-NOM X-ing)
    and
    (NP the fact that Y))
```

11.2 Multiple traces of wh-operators

If there is a strong sense that a relative operator belongs with each of two conjoined S's, it is possible to put in two *T* markers. The example is (irrelevant details omitted):

```
(SBAR (WHPP-1 among
              (WHNP which))
      (S (S (NP-SBJ (NP 139 items)
                    (PP-1 *T*))
            (VP won
                (NP awards)))
         and
         (S (NP-SBJ (NP 11 items)
                    (PP-1 *T*))
            (VP reached
                (NP (ADJP internationally advanced)
                    levels)))))
```

11.3 The "a week before X" pattern

```
(PP-TMP (NP-ADV a week)
        before
        (NP induction))
```

11.4 S-NOM versus Reduced Relative

"With the largest differential occurring in the production of carbaryl methylol" [819#12] is tagged as an S-NOM rather than NP plus reduced relative:

```
(PP with
    (S-NOM (NP-SBJ the largest differential)
           (VP occurring
               (PP in
                   (NP (NP the production)
                       (PP of (NP carbaryl methylol)))))))
```

11.5 In particular

"in particular", "in common", etc. get the following structure

```
(PP in
    (NP particular))
```

11.6 Temperature

We're leaving temperatures flat:

```
(NP 15 degrees C)
```

11.7 Little or no

```
(NP (UCP little or no)
    effect)
```

11.8 Even though

```
(ADVP even
      (SBAR though
            (S X is Y)))
```

11.9 Floating NP

"X inhibits Y, an effect that appears to be Z" [633.17]

An example of NP-ADV at the level of the VP:

```
(S (NP-SBJ X)
   (VP inhibits
       (NP Y)
       (NP-ADV (NP an effect)
               (SBAR (WHNP-1 that)
                     (S (NP-SBJ-1 *T*)
                        (VP appears
                            (S (NP-SBJ-1 *)
                               (VP to
                                   (VP be
                                       (ADJP-PRD Z)))))))))))
```

11.10 Shared S-Level Modifiers

There is no S-level adjunction. Thus an adverbial that modifies both sentences in a coordinated S structure is not adjoined, but rather kept in the form (S (PP X) (S 1) and (S 2)). In cases like this, be very careful that the adverbial really does modify both sentences and can't just be incorporated into the first sentence: this alternative is strongly preferred. If the shared reading is very strong, it is permissible to leave premodifiers at the level of coordination.

```
(S (S (PP-TMP After
              (NP (NP puncture)
                  (PP of (NP (NP coagulated blood)
                             (PP from
                                 (NP the corpora cavernosa))))))
      (NP-SBJ urine retention)
      (VP developed))
   and
   (S (NP-SBJ-1 a suprapubic catheter)
      (VP had
          (S (NP-SBJ-1 *)
             (VP to
                 (VP be
                     (VP introduced
                         (NP-1 *)
                         (ADVP temporarily)
                         (PP-PRP for
                                 (NP urine drainage)))))))))
```

Particularly in sentences with the conjunction "or", premodifiers sometimes clearly modify both sentences, and should be left at coordination level:

```
(S (SBAR-ADV Although X)
   either
   (S Y)
   or
   (S Z))
```

11.11 Flat FRAG for non-body text (by-lines, citations, PMID info, etc.)

FRAG is also used to mark off material that we do not treebank, such as the by-line, citation, and PMID info at the beginning and end of each abstract. This is a fundamentally different use of FRAG than those discussed above, but can be clearly disambiguated from context: this second usage of FRAG is found in sections rather than sentences, and contains no other syntactic nodes inside it. In its standard use, FRAG is not used to bind together arbitrary material; there has to be some relationship between the elements being bound.

This use of FRAG should overlap nearly completely with the use of “section” or SEC as opposed to “sentence” or SENT at the “Sentence” level of annotation, particularly in the BioMedical corpus, where the distinction between SEC and SENT is being made for related entity tagging purposes. “Sentences” are biomedical text, in which we are primarily interested: the title and body of the abstract (regardless of whether they are in fact fully grammatical sentences), and very rarely some other kinds of material. “Sections” include bibliographic information, authors' names and affiliations, PubMed indexing information, copyright notices, and other such extraneous material, even if included in the body text.

11.12 Subheadings

Subheadings, such as "Methods:", "Conclusion:" are tagged as NP and left at top level. They do not get the -HLN function tag, which is reserved for the actual headline (of a news story) or title (or an abstract).

11.13 List markers

List markers are tagged as NP when they fall outside of a sentence, as opposed to the usual LST, which is used when the markers fall inside a sentence.

NP list markers outside of the sentence:

```
(NP Process:)
(NP 1.)
(S The DNA was extracted)
(NP 2.)
(S We tested for the heterozygosity of X)
```

LST list markers contained within the sentence:

```
(PP by
    (NP (NP two independent methods)
        (NP (NP (LST (a) )
                (NP direct sequencing)
                (PP of
                    (NP the PCR product)))
            and
            (NP (LST (b) )
                (NP selective enrichment)
                (PP of
                    (NP mutant DNA))))))
```

11.14 Punctuation addendum (updates to Chapter 3)

We are not applying the paired punctuation rule to commas. (update to 3.1.1.1) Paired commas, as in those setting off an appositive, do not have to be at the same level of the tree. Instead, each comma is put at the lowest possible level, as we do with all punctuation:

```
(S (NP-SBJ (NP (NP St. John's)
               wart)
           ,
           (NP a purported antidepressant))
   ,
   (VP is ...))
```

Punctuation should not be included in constituents being copied or moved, even if it results in incorrect punctuation at the destination site. For example, in "CYP2D6 and CYP3A4 - mediated activities", the hyphen should not be included in the distributed constituent "mediated", even though this results in the reconstructed entities "CYP2D6 mediated" rather than "CYP2D6-mediated."

12 Addendum for other current non-biomedical treebank projects

Changes in Treebank Policy for the English-Chinese Treebank (ECTB) and English-Arabic Treebank (EATB) are the same as those for the biomedical treebank for the most part. Differences between the treebank annotation guidelines for these projects and those presented above for the biomedical project are below.

12.1 Introduction

In addition to the annotation changes instantiated in BioMedical Treebank, several other changes have been applied to the Treebank annotation for the ECTB and EATB projects. Some of the more sweeping changes have been adopted (such as the category NML and the more liberal use of "pseudo-passives"), while others have not (such as the placeholder *P*). In addition, there have been points of departure between the projects in regards to individual items (e.g., "compared with").

This sequel highlights the most consequential changes in policy, some of which are differences with the BioMed project, and others are reiterations of the policy established there. The changes outlined in the sections preceding this one are, in general, to be assumed to apply to the ECTB and EATB corpora. Note that many of changes implemented for the BioMed project simply do not occur in the ECTB and EATB due to the (fairly radical) differences between the corpora both in subject matter and literary style.

12.2 Use of NML

In most instances, the use of the NML tag follows the guidelines prescribed in the preceding sections. For example,

```
(NP the
    (ADJP (NML Hong Kong)
          -
          based)
    company)

(NP-SBJ (NP The number)
        (PP of
            (NP (UCP (NML Chinese-foreign joint ownership)
                      and
                     (ADJP cooperative))
                construction enterprises)))
```

As outlined in 1.1.3 above, in instances where it is difficult to determine the scope of prenominal elements, the default treatment is to group as much as can reasonably be determined and leave the remainder flat,

```
(NP foreign investments and donations)

(NP a foreign enterprise project contracting capital and quality
    license)
```

In some instances, NML is used simply to mark a constituent as a pre-head nominal modifier, regardless of its syntactic category.

```
(NP the
    (NML (S (NP-SBJ (NP a - man 's)
                    -
                    life)
            (VP - is
                (ADJP-PRD - hard))))
    generation)
```

In other instances, it is used to set off a non-nominal constituent that is functioning as an NP head. Note that this and the above usage do not occur in the BioMed corpus.

```
(NP these three
    (NML (S (NP-SBJ *)
            (VP do
                not
                (VP fears))))
```

12.2.1 The use of NML and shared material inside NP

12.2.1.1 Nominal Subconstituents

NML is used to mark nominal subconstituents that do not follow our assumed right-branching default structure:

```
(NP the
    (NML Hong Kong)
    economy)

(NP (NML high level)
    economic talks)

(NP the
    (ADJP (NML New York)
          -
          based)
    company)
```

12.2.1.2 Coordinated Premodifiers

Coordinated premodifiers form a constituent node, typically ADJP, UCP or NML. Following standard policy for coordination, the individual coordinated elements only receive syntactic nodes if one or more of them is multi-token.

```
(NP this
    (NML (NML large scale)
         and
         (NML high level))
    international convention)

(NP (UCP (JJ domestic) and (RB overseas))
    markets)

(NP (UCP (ADJP scientific)
         and
         (ADJP technological)
         and
         (NML software development))
    companies)

(NP the
    (NML (NML Red Cross)
         and
         (NML Red Crescent))
    movement)

(NP his
    (ADJP energetic and powerful)
    performance)
```

12.2.1.3 Coordinated heads with shared premodifiers

In a coordinated NP, any modifiers that are left flat are assumed to be shared across all the heads:

```
(NP China macroscopic economic readjustment and control)

(NP (NP China's)
    international income and expenses)
```

Unshared modifiers in a coordinated structure must form a constituent node with the head they are modifying, as in standard NP coordination:

```
(NP (NP material civilization)
    and
    (NP spiritual culture))
```

When unshared and shared modifiers are combined, the above structure is preserved (using the NML node label) for the unshared components. That is, each of the coordinated elements is marked as a constituent and the elements together form a constituent. Any unshared modifiers are left as sisters to the coordinated structure:

```
(NP socialist
    (NML (NML material civilization)
         and
         (NML spiritual culture)))

(NP (NML New Zealand)
    (NML (NML industries)
         and
         (NML business circles)))
```

12.3 Non-use of *P*

The placeholder *P* (as introduced in 1.2 above) is not used in the ECTB and EATB corpora. Rather, the old Penn Treebank guidelines are followed as much possible. As a result, NML is used rather less in these corpora.

12.4 Changes in Old Tokenization Policy

Certain changes in the old POS policy regarding the tokenization of hyphenated items have resulted in minor changes to treebanking policy. For instance, many adjectives which were previously a single-token under the previous guidelines have been separated into multiple tokens, thus requiring an ADJP node dominating them. For example,

```
(NP (ADJP so - called)
    false cypresses)

(NP a
    (ADJP well - known)
    fact)

(NP (ADJP visa - free)
    travel)
```

Similarly, these changes have created many circumstances requiring the use of NML, e.g.

```
(NP a
    (ADJP (NML Hong Kong)
          -
          based)
          company)
```

```
(NP (NML Taiwan - Palau)
    trade)

(NP (NML (NML Hong Kong)
         -
         (NML Palau))
    trade)
```

12.5 The Use of FRAG in Headlines

In headlines, FRAG is used for a telegraphic/headlinese sentence which is missing a key grammatical constituent (such as determiner, auxiliary, etc.).

```
(FRAG-HLN (NP-SBJ-1 Clinton)
          (S-PRD (NP-SBJ-1 *)
                 (VP to
                     (VP leave
                         (PP-DIR for
                                 (NP Damascus))))))
```

When more than one potential annotation is possible, the current treatment is to include as much annotation as reasonably possible; in other words, to make it look as sentence-like as possible. So, the following is treated as a telegraphic, copula-less sentence rather than a simple NP reduced-relative.

```
(FRAG-HLN (NP-SBJ-1 Three)
          (VP Injured
              (NP-1 *)
              (PP in
                  (NP (NP Clashes)
                      (PP between
                          (NP (NP Israeli Soldiers)
                              and
                              (NP Palestinian Police)))))))
```

12.6 Expanded Use of "Pseudo-passives"

The ECTB and EATB mirror the expanded usage of "pseudo-passives" as described in 3.1 above.

```
(S (NP-SBJ-1 (NP It)
             (SBAR-2 *EXP*)
   (VP was
       (VP reported
           (NP-1 *)
           (SBAR-2 that the joint-investment agreement...))))
```

Slightly more complicated examples (without a passive trace) also occur.

```
(SBAR-TMP soon after
          (S (NP-SBJ four Saudi groups)
             (VP made
                 (S (NP-SBJ (NP *)
                            (SBAR-1 *EXP*))
                    (VP known
                        (SBAR-1 0
                                (S they were setting up…)))))))
```

12.7 Specific decisions

12.7.1 'Compared with' has been given the same treatment as 'compared to', i.e., (*pace* the BioMed treatment described above)

```
(PP compared
    (PP to
        (NP this)))

(PP compared
    (PP with
        (NP this)))
```

12.7.2 'In order to' is now annotated as

```
(PP-PRP in
       (NP order
           (S (NP-SBJ *)
              (VP to ...))))
```

12.7.3 "On board" is treated as a multi-token preposition

```
(NP (NP the two Greek soldiers)
    (PP-LOC on board
            (NP the plane)))
```

12.7.4 For (prescriptively frowned upon) genitive antecedents of relative clauses, we have adopted the following annotation

```
(NP (NP (NP (NP Saed Ibrahim)
            (SBAR-1 *ICH*))
        's )
    position)
    (SBAR-1 who asked for...))
```

12.7.5 RRC occasionally contains -SBJ and -PRD function tags

```
(S (NP-SBJ-2 (NP 401 people)
             (RRC-1 *ICH*))
   (VP have
       (VP been
          (VP killed or injured
              (NP-2 *)
              (PP by
                 (NP-LGS landmines))
              (RRC-1 (NP-SBJ (NP 25%)
                             (PP of
                                 (NP them)))
                     (NP-PRD children)))))
```

12.7.6 Sports and Rankings

12.7.6.1 The -CLR function tag is used on adverbial elements following verbs such as 'came', 'rank' and 'seed' (cf. 5.5 above)

```
(VP came
    (ADVP-CLR in)
    (ADVP-CLR second))

(VP seeded
    (ADVP-CLR fourth))

(VP ranked
    (ADVP-CLR fourth))
```

12.7.6.2 Scores of the format "1-2" are marked as NP-ADV

```
(VP lost
    (NP-ADV (NP 25)
            (PP -
                (NP 28)))
    (PP to
        (NP Qatar)))
```

This is excepted when in prenominal position.

```
(NP the
    (NML (NML 1)
         (PP to
             (NP 1 )))
    tie)
```

12.7.7 "Hold/take hostage" is annotated with a small clause analysis

```
(S (NP-SBJ-1 Three people)
   (VP were
       (VP held/taken
           (S (NP-SBJ-1 *)
              (NP-PRD hostage))))
```