Bracketing Biomedical Text:
An Addendum to Penn Treebank II Guidelines

Colin Warner, Arrick Lanfranchi, Tim O'Gorman,
Amanda Howard, Kevin Gould, Michael Regan

Institute of Cognitive Science, University of Colorado at Boulder

January 2012

Contents

1 Introduction

This document covers the additions and revisions made to Treebank annotation policy in the course of annotating biomedical text, with a particular focus on the unique features of clinical and pathology notes. It is meant to be used alongside the original Penn Treebank guidelines (Bies et al., 1995)[1]. It also relies heavily upon aspects of the Penn Biomedical corpus guidelines (Warner et al., 2004)[7] pertaining to annotation of complex noun phrases, as well as the policy supplements of Treebank 2a (Taylor, 2006)[5] and ETTB 2.0 (Mott et al., 2009)[2], which outline treatment of hyphenated phrases, classes of verbs, and myriad other policy refinements. In addition, speech effects from transcribed notes are annotated according to the Switchboard guidelines (Taylor, 1996)[4] and additional policies have been adopted from the treebank guidelines for the CRAFT biomedical journal article corpus (Xue et al., 2011[8] and Verspoor et al., in press)[6]. In cases of conflicting annotation policies, it is assumed that later policy revisions supersede previous policy.

2 Tokenization

2.1 Sentence-level tokenization

The division between sentence units is determined primarily by the presence of line breaks and final punctuation (. ? !) in the source text.

Bracketed citations following final punctuation are grouped with the preceding sentence as a single sentence unit:

> Studies have shown kava to be an effective therapy, but it can harm the liver.[6][7]

In addition, abbreviations marked with periods (Dr., Fig., M.D., etc) do not trigger a sentence boundary. The list of such abbreviations stands as follows, subject to future additions:

> Adm. Brig. Capt. Cmdr. C.N.P. Co. Col. Cpl. cm. D.O. Dr. Drs. e.g. 'et al.' Fig. Fr. Ft. Gen. Gov. i.e. Lt. lt. Ltd. Maj. M.B.B.S. min. Mr. Mrs. Ms. MI. M.D. No. no. N.P. P.A.-C. Ph.D. P.O. Prof. Prop. Pte. Pvt. Rep. Reps. Rev. R.N. Rt. rt. Sen. Sens. Sgt. St. Ste. tr. U.S. vs.

Due to the presence of initials and the scrambling of titles in de-identified names, we also supplement this list of abbreviations with the following regular expressions:

```
[A-Z][a-z]\.
[A-Z]\.
```

This ensures that sentences such as the following are left as a single sentence unit:

```
Seen with Mp. I. R. Jxpwlwa.
Seen in consultation with Ce. C.J. Ikdgikg.
```

Via these mechanisms, sentence unit divisions can be made deterministically with a high degree of accuracy. However, some degree of annotator correction is still expected when abbreviations fall at the end of a sentence, when encountering previously-unseen abbreviations that do not fall under existing patterns, or when the formatting of the source text has inserted extraneous mid-sentence newlines.

2.2 Word-level Tokenization

White space in the source text is always considered to mark a token boundary. Additional token boundaries are defined in accordance with Mott et al., 2009[2], Section 1, with the following revisions:

2.2.1 Hyphenated Words

A number of biomedical affixes have been added to the list of hyphenated interjections and affixes that are kept as single tokens. The list now stands as follows:

> adeno- aorto- ante- anti- arch- ambi- -able -ahol -aholic -ation axio- be- bi- bio- broncho- co- counter- cross- centi- circum- cis- colo- contra- cortico- cran- crypto- -centric -cracy -crat de- deca- demi- dis- -dom eco- electro- ennea- -esque -ette ex- extra- -er -ery ferro- -ful -fest -fold gastro- -gate -gon giga- hepta- hemi- hypo- hexa- -hood in- inter- intra- -ian -ible -ing -isation -ise -ising -ism -ist -itis -ization -ize -izing ideo- idio- infra- iso- -less -logist -logy -ly judeo- macro- micro- mid- mini- mono- musculo- neuro- nitro- mm-hm mm-mm -most multi- medi- milli- neo- non- novem- octa- octo- o-kay -o-torium ortho- over- paleo- para- pelvi- pheno- penta- pica- poly- preter- pan- peri- phospho- pneumo- pseudo- post- pre- pro- quasi- quadri- quinque- -rama re- recto- salpingo- sero- semi- sept- soci- sub- super- supra- sur- tele- tera- tetra- tri- uber- uh-huh uh-oh ultra- un- uni- vice- veno- ventriculo- -wise x-ray x-rays x-rayed

Note that this list is case-insensitive and non-exhaustive. Future additions are likely as additional affixes are encountered.

Examples of unsplit hyphenated phrses:

```
non-pharmacologic
anti-inflammatory
```

```
re-evaluation
anti-anxiety
```

All other (non-numeric) hyphenated phrases that do not contain an affix from the above list are split into multiple tokens.

2.2.2 Abbreviations Containing Slashes

Abbreviations containing slashes remain a single token when the full abbreviations resolve to a single word: (list is comprehensive for current data)

```
b/c   (for 'because')
d/c   (for 'discontinue')
u/s   (for 'ultrasound')
w/o   (for 'without')
```

However, abbreviations resolving to multiple words are split into multiple tokens. Representative examples include:

```
c / o   (for 'complain of')
n / a   (for 'not available')
f / u   (for 'follow up')
s / p   (for 'status post')
y / o   (for 'year old')
```

2.2.3 Dates and Phone Numbers

Dates are kept as single tokens when represented with slashes or hyphens.

```
12-5-95
1995-3-1
5/4/98
6/15
```

This includes de-identified dates that do not necessarily fall within standard numerical ranges.

```
2883-5-26
7552-3-73
25/16
25/26/0551
```

Phone numbers and other hyphenated numerical identifiers (e.g., grant numbers or patient numbers) are left as single tokens.

```
9646-66
9-2282
968-9060
```

2.2.4 Ranges and Proportions

Ranges of numbers containing dashes are split into multiple tokens.

```
days 16 - 25
70 - 99 percent
```

Slashes used in numerical ratios are also split.

```
blood pressure reading of 220 / 100
grade 4 / 4
```

Both of these patterns can potentially resemble phone numbers or dates, which are not split. Thus, this is one of the few areas where annotator correction of automatic tokenizer output is sometimes required. This is especially true if the source text is encoded in low level ascii and does not make a typographic distinction between hyphens and dashes.

2.2.5 URLs and Email Addresses

URLs, e-mail addresses, and file names are kept as single tokens.

```
http://seer.cancer.gov/csr/1975_2005/index.html
info@ndif.org
```

2.2.6 Other Hyphens, Dashes, and Slashes

Apart from the patterns mentioned above, all other hyphens, dashes, and slashes are split into separate tokens.

```
salt - free
94 - year - old
1.20 L / h / kg
and / or
Tonsillectomy / appendectomy
```

2.2.7 Units

The following units are split off into separate tokens when appended to numerals:

> am cc cm d days Fr kg km l m mcg mg min minutes ml mm mos pm units y yr yrs

For example, '3cm' is split into the two tokens '3' and 'cm'. This list is case-insensitive and non-exhaustive. It represents units encountered to date.

2.2.8 Other Punctuation and Symbols

With the exception of the patterns outlined above, the following symbols and punctuation are split:

> , : ; . @ # $ % / + - < > ∼ ≤ ≥ ± × ÷

The pattern 3°C has been standardized to two tokens: '3' and '°C'

2.2.9 Abbreviations

Alphanumeric abbreviations that resolve to multiple words are left as one token.

```
degC (for 'degrees Celsius')
DNR (for 'do not resuscitate')
NKDA (for 'no known drug allergies')
```

3 Part of Speech Tags

3.1 Overview of Tagset

> VB VBP VBZ VBD VBN VBG NN NNS NNP NNPS JJ JJR JJS RB RBR RBS WP WP$ WRB WDT PRP PRP$ DT PDT IN TO MD RP POS CD CC EX . , : `` '' -LRB- -RRB- SYM $ LS UH FW
>
> HYPH AFX GW XX

Most of these tags are described in the Part of Speech Tagging Guidelins (Santorini, 1990)[3]. The final four tags HYPH, AFX, GW, and XX are covered in subsequent guideline supplements. HYPH (unbound hyphen) and AFX (unbound affix) are described in Warner et al., 2004[7], GW (mistranscription) in Taylor, 1996[4], and XX (uninterpretable material) in Mott et al., 2009[2].

3.2 TO versus IN

TO is used for infinitival 'to' only. Prepositional 'to' is tagged IN.

3.3 HYPH

Split hyphens receive the HYPH tag.

```
94 -/HYPH year -/HYPH old
sleep -/HYPH related
```

Unpronounced slashes also receive the HYPH tag.

```
and //HYPH or
c //HYPH o (complains of)
s //HYPH p (status post)
1 //HYPH 2 tablet
```

3.4 AFX

Affixes occurring on their own receive the AFX tag. This includes not only affixes from the list in section 2.2.1 above, but also stranded morphological units like plural 's'.

```
pre-/AFX and postoperatively
non/AFX lung cancer
mid/AFX 20s
location ( s/AFX )
factor ( s/AFX )
```

3.5 GW

The GW (Goes With) tag is used to reconstruct a single lexical item that has been erroneously written as multiple tokens (ie, a word with medial whitespace in the source text). The final token has the correct POS tag and all tokens to be joined to it are tagged GW.

For example, "no where" in the source text is tagged as:

```
no/GW where/RB
```

Other examples:

```
bis/GW ected/JJ
may/GW be/RB   (for 'maybe')
```

3.6 XX and De-identified Data

If a de-identified phrase is clearly a proper name within the context of the larger sentence, it is POS-tagged as NNP:

```
Dn./NNP Xzfhqc/NNP is a 33 - year - old white male .
El./NNP Wr/NNP Ezlzl/NNP is a 90 yo gentleman from Jeizc/NNP , WN/NNP .
```

Other, less interpretable de-identified phrases are POS-tagged XX:

```
He recently jpcfjpu/XX njme/XX Bpwjq/XX wq/XX w/XX qwbpqewg/XX
He dzyedzu/XX in 1972
who has previously worked as a dstyx/XX , and inzgsu/XX pnxvyx/XX
using his hands quite a bit for voxhdpyxg/XX opk/XX nopktvohqpb/XX
```

3.7 Use of SYM

3.7.1 Non-alphanumeric List Markers

```
-/SYM Topical Steroid Therapy
o/SYM Spicy Foods
*/SYM Infertility
"/SYM Diabetes
@/SYM adults
```

(These reflect corpus-specific typographic conventions for handling headings and lists)

3.7.2 Pronounced Punctuation

'-' is tagged as SYM when pronounced, typically as 'to' or 'negative'

```
-/SYM 1 to -/SYM 2
ages 92 -/SYM 82
2 -/SYM 3 days
```

'/' is tagged as SYM when pronounced (eg, 'per', 'of' or 'over')

```
60 //SYM minute
130 mg //SYM dL
```

```
180 //SYM 90 mmHg
```

'x' is tagged as SYM when pronounced, typically as 'by' or 'times'

```
15 x/SYM 10 x/SYM 3 cm
qd x/SYM 1 week
4 x/SYM a week
```

':' is tagged as SYM when pronounced, typically as 'to'

```
ratio of 6 :/SYM 1
```

'?' is tagged SYM when it is not final punctation, but functioning with the meaning of 'possibly' or 'unknown' or to indicate author uncertainty.

```
a dose of ?/SYM for treatment
?/SYM secondary to Kenalog Orabase
some small ?/SYM ulcerations
reading was 170 / ?/SYM .
```

3.7.3 Mathematical Symbols

Mathematical symbols such as the following are POS-tagged SYM:

\+ - / * = < > ± ≤ ≥ ~ @

3.8 Other Symbols

```
#/NN
%/NN
'/POS   (possessive plural)
'/NN(S) (meaning foot/feet)
"/NN(S) (meaning inch/inches)
&/CC
$/$     (and other currency terms such as US$/$ USD/$ RMB/$)
°/NN(S) (and other temperature terms such as °F/NN(S) and °C/NN(S))
```

3.9 FW

Dosage abbreviations derived from Latin are POS tagged FW. Examples include:

```
bid/FW (from Latin 'bis in die', meaning 'twice a day')
p.r.n./FW (from Latin 'pro re nata', meaning 'as needed for')
```

```
q.h.s./FW
p.o./FW
pr/FW
qd/FW
```

3.10 List Markers

List markers such as '#1.', 'A.', and 'A)' are POS-tagged as follows:

```
#/NN 1/LS ./:
A/LS ./:
A/LS )/-RRB-
```

3.11 NNP versus NN

For the sake of annotation consistency, capitalization is the key feature used to distinguish between common nouns (NN/NNS) and proper nouns (NNP/NNPS). One exception is genus-species collocations, in which the lower-case species names is tagged NNP:

```
E./NNP coli/NNP
Pneumocystis/NNP carinii/NNP
```

3.12 Abbreviations

As a general rule, an abbreviation receives the same POS tag as its expanded form.

3.12.1 NN versus NNS

Abbreviated units can be tagged as NN or NNS, depending on the expanded form in context would be singular or plural:

```
a 3.3 cm/NN cyst
a 120 mg/NN tablet
40 mg/NNS by mouth
approximately 1.25 cm/NNS in diameter
164.0 cm/NNS ( 64.5 in/NNS )
```

Abbreviations that resolve to multi-token phrases receive the POS tag of the head of that phrase:

```
3 cfn/NNS (3 centimeters/NNS from nipple)
68 degF/NNS (68 degrees/NNS Farenheit)
```

3.12.2 NN versus NNP

Abbreviations resolve to NN versus NNP depending on the conventional capitalization of the expanded form:

```
UTI/NN  (for 'urinary tract infection')
HSP/NNP (for 'Henoch-Schonlein Purpura')
```

3.13 List Markers

4 Treebank

4.1 Overview of Tagset

4.1.1 Node Labels

S SBAR SBARQ SQ SINV WHNP WHPP WHADVP WHADJP NP VP PP ADVP ADJP LST NML PRN PRT QP FRAG CONJP INTJ NAC RRC UCP EDITED X

4.1.2 Function Tags

-SBJ -PRD -DTV -LGS -VOC -SEZ -IMP -ETC -NOM -ADV -BNF -DIR -EXT -LOC -MNR -PRP -TMP -TPC -CLR -CLF -TTL -UNF -RED

Note that the -HLN tag (used to mark headlines in newswire articles) is inapplicable to biomedical texts and has been dropped from the current tagset.

4.1.3 Empty Categories

* *PRO* *T* *RNR* *ICH* *EXP* *U* 0 *?* *NOT*

Usage of most empty categories is explained in the original Treebank II guidelines[1]. The Treebank IIa guidelines[5] introduce *PRO* and cover the difference between *PRO* and *. Note that the empty category *P* from the Penn Biomedical guidelines[7] is not used.

4.2 -RED

Fragmentary sentences that are missing copulas or auxiliaries appear with a high degree of frequency in the abbreviated style of clinical notes. This has prompted the refinement of treebank policy for these patterns. A new function tag -RED (REDuced) has been introduced to indicate that a sentence is missing a copula or auxiliary. Every S-RED should have a subject (*-SBJ) and a predicate (VP or *-PRD). If there is no overt subject, an arbitrary subject (NP-SBJ *PRO*) is inserted.

Under prior treebank policy guidelines, clauses missing a copula or auxiliary were analyzed as FRAG and did not consistently receive -SBJ and -PRD labels or empty subjects. With S-RED, more argument structure is provided for the abbreviated patterns found in clinical notes, and FRAG is restricted to material that does not contain predicate argument structure.

4.2.1 Non-Verbal Predicates Missing Copulas

Sentences that are missing a copula are analyzed as S-RED with a non-verbal predicate.

- NP-PRD

```
(S-RED (NP-SBJ Total cholesterol)
       (NP-PRD
           (QP approximately 220))
       .)
```

(Reading: "Total cholesterol [is] approximately 220".)

- ADJP-PRD

```
(S-RED (NP-SBJ Status)
       (ADJP-PRD indeterminate)
       .)
```

(Reading: "Status [is] indeterminate".)

- PP-PRD

```
(S-RED (NP-SBJ Elderly patient)
       (PP-LOC-PRD in
                   (NP care center))
       (PP with
           (NP cough))
```

```
.)
```

(Reading: "Elderly patient [is] in care center with cough".)

4.2.2 Verbal Predicates Missing Auxiliaries

Sentences that are missing an auxiliary are analyzed as S-RED.

• Past Participle

```
(S-RED (NP-SBJ-1 Patient)
       not
       (VP seen
           (NP-1 *))
       .)
```

(Reading: "Patient [was] not seen".)

• Present Participle

```
(S-RED (NP-SBJ Patient)
       (VP having
           (NP significant hot flashes))
       .)
```

(Reading: "Patient [is] having significant hot flashes".)

4.2.3 Bare Adjective or Participle

Sentences consisting solely of an adjective or past or present participle are analyzed as S-RED with arbitrary subject (NP-SBJ *PRO*):

• Bare adjective

```
(S-RED (NP-SBJ *PRO*)
       (ADJP-PRD Obese)
       .)
```

(Reading: "[Patient is] obese".)

• Bare present participle

```
(S-RED (NP-SBJ *PRO*))
       (VP Coughing
```

```
            (PRT up)
            (NP purulent material))
        .)
```

(Reading: "[Patient is] coughing up purulent material".)

• Bare past participle

```
(S-RED (NP-SBJ-1 *PRO*)
       (VP Seen
           (NP-1 *)
           (NP-TMP 2/18/2001))
       .)
```

(Reading: "[Patient was] seen 2/18/2001".)

This means that ADJP and VP are never seen as the top-level node in a sentence. Bare ADJP always projects an S-RED and takes an arbitrary subject. Bare VPs headed by participles project S-RED, bare VPs headed by other verbal forms project a pro-drop S or imperative S-IMP.

4.2.4 S-RED versus NP

If it is ambiguous from context and surrounding syntax whether a statement should be analyzed as an S-RED or an NP, it is annotated as an S-RED with subject and predicate rather than as a single NP containing a reduced relative or adjunct.

```
(S-RED (NP-SBJ-1 (ADJP (NP 30 - year)
                       - old)
                 female)
       (VP exposed
           (NP-1 *)
           (PP to (NP influenza))
       .)

(S-RED (NP-SBJ Patient)
       (VP going
           (PP-DIR to
                   (NP Viet Nam)))
       .)
```

4.2.5 Coordinated S-RED

Coordinated S-REDs take S as their parent node.

```
(S (S-RED (NP-SBJ *PRO*)
          (VP feeling
              (AJDP-PRD flushed)))
   and
   (S-RED (NP-SBJ hands)
          (VP sweating))
   .)
```

Likewise, S coordinated with S-RED takes S as the parent node.

```
(S (S-RED (NP-SBJ-1 Prescription)
          (VP given
              (NP-1 *)
              (PP-DTV to
                      (NP the patient))))
    and
    (S (NP-SBJ-2 She)
       (VP was
           (VP referred
               (NP-2 *)
               (PP to
                   (NP Zmackglkuo & Ymkgakuo Mccaoutklc)))))
   .)
```

4.3 Top Level Nodes

4.3.1 Invalid Nodes

The node labels that are explicitly prohibited from being top-level nodes are ADJP, VP, WHNP, WHADJP, WHADVP, NML.

• Bare ADJP and VP always project an S. (see 4.2.3)

• Bare WHNP, WHADJP, and WHADVP always project a FRAG. (see 4.5.1)

• The NML node label only occurs inside of NP. (see Penn Biomed guidelines[7], section 12.2.1)

• Top level parentheses do not receive a PRN node. Rather, the parentheses are included

in the constituent they surround. For example, the sentence '(IRB number is 06-978234).' is treebanked as top-level S.

```
(S (
   (NP-SBJ IRB number)
   (VP is
       (NP-PRD 06-978234))
   )
   .)
```

4.3.2 Valid Nodes

NP, PP, and ADVP are permitted as top-level nodes. They do not project an S-RED or FRAG, but can stand on their own as the top-level syntactic node of a sentence unit.

```
(NP (NP (ADJP (NP 3 - year)
              - old)
        boy)
  (PP with
      (NP dysuria))
  .)

(PP On
    (NP Premarin, Provera, and Prozac)
    .)

(ADVP Sometimes
      .)
```

4.4 Pro-drop

Pro-drop sentences are more common in the abbreviated language of clinical notes than in many other genres of written English. Pro-dropped sentences are annotated as normal sentences S with an empty subject (NP-SBJ *PRO*).

```
(S (NP-SBJ *PRO*)
   (VP complains
       (PP of
           (NP nausea)))
   .)
```

4.5 FRAG

The FRAG node label is used only for material that cannot be analyzed as forming a standard syntactic phrase like S or NP. A FRAG does not contain predicate argument structure. It joins material which is related, but not by a standard syntactic relationship like SBJ/PRD, coordination, or apposition.

4.5.1 WH- nodes outside of SBAR

Outisde of SBAR, WH- nodes project FRAG. This applies both to bare WH- nodes and WH nodes combined with adjuncts.

```
(FRAG (WHNP how much)
      ?)

(FRAG (WHADVP why)
      ?)

(FRAG (WHNP what)
      (PP about (NP diabetes))
      ?)
```

4.5.2 Conjunction + XP

Sentences consisting of a conjunction plus (non-clausal) constituent are annotated as FRAG.

```
(FRAG But
      (NP (QP about 900,000) platelets)
      .)
```

4.5.3 Headings

FRAG is commonly used to join a heading with the material that follows.

```
(FRAG (NP Discussion and recommendations)
      :
      (S (LST ( 1 ) )
         (NP-SBJ We)
         (VP discussed
            (NP the Registry
                (NML objectives and procedures))))
```

```
      .)

(FRAG (NP Axis II)
      :
      (NP No diagnosis)
      .)

(FRAG (NP HEENT)
      :
      (S (NP-SBJ *PRO*)
         (VP wears
             (NP dentures)))
      .)

(FRAG (NP General)
      :
      (S-RED (NP-SBJ *PRO*)
             (ADJP-PRD (ADJP Alert)
                       ,
                       (ADJP well kept)
                       ,
                       (ADJP oriented)))
      .)

(FRAG (NP Morphone)
      -
      (NP rash))
```

But note that headings can form the subject of an S-RED when semantically appropriate. The default annotation for sentence units that start with headings is still FRAG: the S-RED is only used with a small number of headings that always take predicates.

```
(S-RED (NP-SBJ Home phone #)
       :
       (NP-PRD 318-4396)
       .)

(S-RED (NP-SBJ Referring Physician)
       :
       (NP-PRD Secxd Y. Tecei))
```

4.5.4 De-indentified text

Uninterpretable de-identified text that has received the POS tag XX is treebanked as a flat FRAG:

```
(S (NP-SBJ He)
    (ADVP-TMP recently)
    (FRAG jpcfjpu/XX njme/XX Bpwjq/XX wq/XX w/XX qwbpqewg/XX))
```

But note that de-identified material that has been determined to be nominal (see section 3.6) is treebanked as NP:

```
(S (NP-SBJ Dn./NNP Xzfhqc/NNP)
   (VP is
       (NP a
           (ADJP (NP 33 - year)
                 - old)
           white male))
   .)
```

4.6 X

X is used to 'X out' tokens that do not belong in a sentence. These tokens may be duplicated or otherwise mistyped words, or stray metadata from improperly formatted source files:

```
(NP (NP (QP one - to - two)
        tablets)
    (ADVP orally)
    (X every)
    (NP-TMP every
            (QP four to six)
            hours))

(NP (NP a small breakfast)
    and
    (X a)
    (NP (ADVP-TMP usually)
        a late lunch)

(FRAG (X </ref>) [ (NP 12) ])
```

4.7 Medication and Dosages

Constituents that describe dosage timing or method are adjoined to the dosage amount as adjuncts:

```
(NP (NP 20 mg)
    (NP-TMP a day))

(NP (NP 200 mg)
    (PP by
           (NP mouth))
    (PP-TMP before
           (NP bedtime)))

(NP (NP 250 mg)
    (NP-TMP (NP four times)
            (NP-TMP a day)))

(NP (NP 40 mg)
    (ADVP p.o.)
    (ADVP-TMP tid))

(NP (NP 50 mg)
    (ADVP-TMP a.m.))

(NP (NP 50 mg)
    (ADVP-TMP q.h.s.))
```

The combination of drug name and dosage is analyzed as a single nominal constituent. In the following example sentences, these 'drug plus dose' constituents are marked by parentheses.

> Will add (Colace 100 mg by mouth three times a day).
>
> A prescription for (Synthroid 0.112 mg daily) was provided.
>
> He is on (Synthroid 0.125mg daily) at this juncture.
>
> A year ago, she was started on (Risperdal 0.5 mg twice a day) because of severe anxiety.

We treat the dosage information as an NP adjunct (NP-ADV) adjoined to the drug name.

```
(NP (NP Colace)
    (NP-ADV (NP 100 mg)
            (PP by
```

```
                (NP mouth))
            (NP-TMP (NP three times)
                    (NP-TMP a day))))

(NP (NP CellCept)
    (NP-ADV (NP (NP 50 mg)
                (ADVP-TMP a.m.))
            ;
            (NP (NP 1 g)
                (ADVP-TMP p.m.)))
```

4.8 NP-ADV adjuncts within NP

NP-ADV adjuncts to NP are uncommon in other domains, but very common in the abbreviated style of clinical notes. Aside from the dosage patterns outlined above, these NP-ADV adjuncts also occur frequently in pathology reports, in descriptions of samples taken or areas examined:

```
(NP (NP 7 cm)
    (NP-ADV ileum))

(NP (NP 11 cm)
    (NP-ADV small bowel))

(NP (NP (NP (NML 3.0 cm)
            segment)
        (NP-ADV terminal ileum))
    , and
    (NP (NP 25 cm)
        (NP-ADV (NP cecum)
                and
                (NP right colon))))
```

In the first example, the phrase '7 cm ileum' refers to a 7 cm piece of the (much larger) ileum. The alternative tree (NP (NML 7 cm) ileum) refers to a 7 cm ileum, which is inaccurate.

4.9 -CLR

Following the policy established in the CRAFT biomedical guidelines [8], the -CLR function tag is not applied to PP. S-CLR is still used to label secondary predicates and resultatives, as described in the Treebank 2a guidelines [5].

References

[1] Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. *Bracketing Guidelines for the Treebank II-style Penn Treebank Project*. University of Pennsylvania, 1995. http://www.ircs.upenn.edu/arabic/manuals/root.pdf.

[2] Justin Mott, Ann Bies, Colin Warner, and Ann Taylor. *Supplementary Guidelines for ETTB 2.0*. University of Pennsylvania, 2009. http://www.seas.upenn.edu/~jmott/2009_addendum.pdf.

[3] Beatrice Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing)*, 1990. http://www.cis.upenn.edu/~bies/manuals/tagguide.pdf.

[4] Ann Taylor. *Bracketing Switchboard: An Addendum to the Treebank II Guidelines*, 1996. http://www.seas.upenn.edu/~jmott/prsguid2.pdf.

[5] Ann Taylor. *Treebank 2a Guidelines*, 2006. http://www-users.york.ac.uk/~lang22/TB2a_Guidelines.htm.

[6] K. Verspoor*, K.B. Cohen*, A. Lanfranchi, C. Warner, H.L. Johnson, C. Roeder, J.D. Choi, C. Funk, Y. Malenkiy, M. Eckert, N. Xue, W.A. Baumgartner Jr., M. Bada, M. Palmer, and L.E. Hunter. *A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools.* BMC Bioinformatics, in press.

[7] Colin Warner, Ann Bies, Christine Brisson, and Justin Mott. *Addendum to the Penn Treebank II Style bracketing Guidelines: BioMedical Treebank Annotation*. University of Pennsylvania, 2004. http://www.cis.upenn.edu/~bies/bioie/TBguidelines-addendum.pdf.

[8] Nianwen Xue, Arrick Lanfranchi, Colin Warner, Amanda Howard, and Tim O'Gorman. *CRAFT addendum to PTB II and Penn BioIE guidelines*, 2011. http://bionlp-corpora.sourceforge.net/CRAFT/CRAFT_syntax_annotation_guidelines.pdf.