Bracketing Webtext: An Addendum to Penn Treebank II Guidelines

Justin Mott, Ann Bies, John Laury, Colin Warner
Linguistic Data Consortium
{jmott|bies|jlaury|colinw}@ldc.upenn.edu

January 2012

Contents

1 Introduction

The guidelines followed here are based on the original Penn Treebank guidelines (Bies et al., 1995)[1] with speech and similar effects annotated as per the Switchboard guidelines (Taylor, 1996)[3]. In addition, it follows important aspects of the Penn BioMedical corpus (Warner et al., 2004)[4] (however, the placeholder *P* described there has *not* been adopted) as well as the supplementary guidelines described in (Mott et al., 2009)[2], which includes a summary of the 'Penn Treebank IIa' guidelines. It should be noted that later

policy revisions supersede previous policy.

Webtext poses many interesting challenges which do not occur in previously Treebanked material. Among them are the following features.

•Emoticons

```
;) :-) ;) :P
```

•Text Decoration

```
|--------+----------------------->
|        |          Tana.Jones@en|
|        |          ron.com      |
|        |                       |
|        |          11/29/2000   |
|        |          04:34 PM     |
|        |                       |
|--------+----------------------->


  /\___/\
 ( o   o )
 (  =^=  )
 (        )
 (         )
 (  Tana    ))))))))))))
```

•Semi-conventional Web Spellings

srsly, needz, moar, kitteh, hawt, c, u, ur, yr, teh, plz, k, thx, ghey, <3, l8r, gr8, 4, 4evah, 2, whatevs, totes, probs, etc.

(seriously, needs, more, kitty, hot, see, you, your/you're, your/you're, the, please, okay, thanks, gay, love/heart, later, great, for, forever, to/too, whatever, probably)

•Web Initialisms and Collapsed Phrases

idk, 'sup, ftw, lol, btw, imho, kthxbai, g2g, syl, omg, etc.

(I don't know, what's up, for the win, by the way, in my humble opinion, okay thanks goodbye, got to go, see you later, oh my god)

•Non-conventional Spacing

omg i <3 it :-) srsly its the B E S T . B U R R I T O . E V E R !!!

2 Tokenization

2.1 Sentence-Level Tokenization

Previous Treebanks created at LDC have used as their source output of of either careful human editing or translation. As such, line breaks in the source data provided a reliable starting point for syntactic annotation. Because of this, an almost entirely automated system was used to determine the boundaries between sentences. However, these deterministic methods of splitting sentences are not wholly adequate for web some text data for at least two reasons. First, the line breaks extant in the source material are unconstrained and occur for any number of formatting or other reasons, e.g.,

> The U.S. and Europe are constantly improving their defense against the
> last
>
> murder, not the next one. We may arrange for the best airport
> security in the world.

Human readers, though, realize immediately that the correct parsing of those sentences is as follows.

> (S The U.S. and Europe are constantly improving their defense against the last
> murder, not the next one.)
>
> (S We may arrange for the best airport security in the world.)

Second, line breaks are often inserted for stylistic reasons, such as in the following sentence, where it is clear that the entire postal address is the predicate of the verb 'is'. Strictly following the existing line breaks would greatly hamper the accurate syntactic annotation of the sentence by stranding the predicate from the verb. This type of text is much more pervasive in web text than in previously-Treebanked material. For web text data with such unreliable line breaks the guidelines below are applied. For web text with reliable line breaks, normal sentence tokenization procedures are applied.

> The new postal address is:
> Global Campaign for Education,
> PO Box 521733,
> Saxonworld,
> Johannesburg 2132,
> South Africa

Conversely, as in the original Penn Treebank, it is possible to identify specific patterns that can occur within the same line, but do not have a syntactic relationship with surrounding text. In the Penn Treebank, these were restricted to headlines, bylines, datelines, etc., due the style of text being analyzed. There are a number of additional types of easily-identifiable patterns that occur frequently in webtext. Identifying and splitting these patterns should improve the quality of the Treebank annotation.

For these reasons, it is now necessary to allow human annotators to correct the output of automated sentence splitters in order to rectify these problems. The guidelines outlined below are an attempt to maximize the amount of syntactic information that can be represented in the Treebank while keeping the sentence-level annotation as straightforward as possible. To that end, line breaks existing in the source texts are to be followed as much as possible.

2.1.1 General Guidelines

Token strings are split into separate sentence units in the following situations. Note that the text is presented as it appears in the source text and sentence units are represented with the notation [SU].

When final punctuation is used at the end of a sentence. Final punctuation is defined as a period, question mark, or exclamation point. In practice, this is done using an automatic sentence splitter.

> [SU John can come over to eat at my house anytime.] [SU -Juggernaut]

Emoticons and stray punctuation are joined with the preceding sentence.

> [SU Thanks! :-)] [SU See you later!]

Colons do not count as final punctuation:

> [SU This is hardly surprising: local authorities and administrations have still to learn how to cope with the disaster caused by the tsunami.]

Note that final punctuation does not cause a split when it exists within a single token.

> A.D.
> www.wikipedia.com
> Yahoo!

Final punctuation does not cause a split when it is used in error.

> [SU John Lennon's divorce followed, as well as his entering the world of black magic, as deeply as to buy the apartment. where the "Rosemary's baby" had

been filmed, previous property of Roman Polansky, and in that same apartment John Lennon had a room upholstered with black silk where he used to do his black magic operations.]

Ellipsis is used regularly as final and non-final punctuation in English writing. Therefore, ellipses trigger sentence splits when functioning as a period and does not otherwise. In cases where its usage is ambiguous, it will be treated as medial punctuation.

[SU Come and say hello...you will be warmly welcomed!]

2.1.2 Specific Functionally-Defined Patterns

1. Title/heading/headline

   [SU GCE ASKS TANZANIA GOVERNMENT TO RECONSIDER BAN OF EDUCATION NGO, HAIKIELIMU] [SU In response to the government of Tanzania's ban on research by the NGO, HakiElimu, GCE has written directly to the Minister of Education, asking him to reconsider.]

2. Dateline (and similar lines for author/place/source information)

   [SU Boulder, CO, Feb. 23 (UPI) –] [SU Ever since he ate up Red Riding Hood's grandma and blew down the houses of two-thirds of the Three Little Pigs, the Big Bad Wolf has held a persistently bad reputation.]

   [SU (BBC)] [SU The Palestinian militant organisation Hamas has announced an end to rocket attacks on Israel from the Gaza Strip after a weekend of escalating violence.]

3. Epistular salutation

   [SU Roy,] [SU I am sending you some info that I have passed along regarding the effort to win the cash for the fund raising benefit.]

4. Epistular valediction

   [SU kisses] [SU leili]

5. Signature

   [SU I am still on earth.. because there's a goddess here who loves me more"] [SU C Mayur Shah]

2.1.3 Joining Lines into Sentence Units

As stated above, chunks of text with internal line breaks in the source data are now being joined into sentence units when there is syntactic structure to preserve. Some examples include:

1. Argument of a verb

   [SU You will discover:
   •How to Create the Security of Residual Income;
   •Why Agel is the Best Vehicle to Live Your Dreams;
   •The Amazing Scientific Breakthrough Behind the Products;
   •How the "Quadra-Plan" Builds Bigger Bonus Checks Faster;
   •The Support System in Place to Help You Build Stronger; and
   •The Secret to Lock in a "Legacy" Position!]

2. Argument of a preposition

   [SU COMPLETE ARTICLE AT:
   http://discountairlineticket.blogspot.com/...]

3. Sentential adjunction

   [SU So to summarise :-
   Man 1 can see men 2 and 3.]

   [SU Man 2 can see man 3.]

   [SU Man 3 can see none of the others.]

   [SU Man 4 can see none of the others.]

4. Other

   [SU After every world war, the rules of international

   law have changed, and the same will happen after the present one.]

When there is ambiguity as to whether there is a (non-linguistic) header-list relationship or a (linguistic) appositive relationship, the former is assumed for simplicity's sake.

[SU Mail the six envelopes to the following addresses:]

[SU 1) G. Burrows
1/264 Tor St
Toowoomba QLD
4350 Australia]

[SU 2)S Luest
P.O. Box 366
Garden Grove, CA 92842
US.]

As mentioned above, postal addresses and other blocks of contact information are joined into sentence units.

[SU Tom Dempsey
Gulf Breeze, FL
850-748-0740]

[SU Animals R Us.Net
Bill Schmidt
animalsr...@aol.com
732-657-3416
1027 Jolson Court
Manchester, N.J. 08759]

[SU Email : mayur...@yahoo.com
SMS : +919819602175
Web :]

2.2 Word-Level Tokenization

White space is *never* deleted at the word level, except cases where new lines have been removed in the sentence tokenization as described above. The following examples were split in the source and were left split throughout the annotation process.

```
basic ally (for ‘basically’)
mellow dramatic (for ‘melodramatic’)
realtion ship (for ‘relationship’)
```

Spelling errors interfering with automated tokenization processing were corrected in the direction of allowing correct POS and TB annotation. In the following examples, the output of the automated token script was changed as indicated in the round brackets.

```
the scammer s highest income  (scammers -> scammer s)
their mezzo luna’s are deffly better (luna ’s -> luna’s)
they’re feet need to be done (they ’re -> they’re)
thei r energetic at night (their -> thei r)
the re from iraland (there -> the re)
```

Obvious finger-slip errors which triggered splitting by the automated tokenizer were corrected. In the first two examples, 'p' and 'm' were replaced by '[' and ',' respectively, one key to the right of the intended target. In the last example, presumably the writer attempted to delete the 'h' and hit '=':one key left of the delete key.

```
ha[[y (for 'happy'; ha [ [ y- > ha[[y)
I ', (for 'I'm'; I ' , -> I ',)
h=guys (for 'guys; h = guys -> h=guys)
```

The following are split as below.

```
male(s) -> male ( s )
outta -> out ta
dunno -> du n no
```

White space is inserted around slashes if the constituent parts resolve to words.

```
a/c -> a / c (for 'air conditioner')
b/c -> b/c (for 'because')
```

Initialisms and abbreviations, however, are not broken down into constituent tokens, even though this can interfere with POS and TB annotation.

```
idk where to go (for 'I don't know')
atm (for 'at the moment')
roflmao (for 'rolling on the floor, laughing my ass off')
```

At-signs are separated off as well. No hash tags were encountered in this data set, but would be split off if seen.

```
@ W.a.b.b.y
```

3 POS Tagging

3.1 Typos, Misspellings, etc.

Words are assigned their correct POS tags regardless of spelling anomalies.

```
they're/PRP$ feet need to be done
thei/PRP r/VBP energetic at night
the/PRP re/VBP from iraland
```

Obvious finger slips are treated likewise as well.

```
h=guys/NNS
```

```
you just can't be ha[[y/JJ
I ',/VBP a mormon
```

3.2 Use of GW

As in the Switchboard guidelines, words broken up by added white space are tagged as follow: GW is used for the non-final tokens and the correct tag is placed on the final token. This holds for anomalous uses of hyphens, unintentional white space and intentional white space. Some examples are:

```
basic/GW ally/RB
mellow/GW dramatic/JJ
realtion/GW ship/NN
```

It is also used to hold together constituent parts of file names containing white space with no linguistic meaning, as well as email addresses containing white space.

```
ENRON/GW -/GW Para13/GW -LRB-/GW nmemdrft8-7-01/GW -RRB-/GW .doc/NN
Billy/GW Dorsey@ENRON_DEVELOPMENT/ADD
```

3.3 Use of NFP (Non-Final Punctuation)

3.3.1 Emoticons

The NFP tag is used for emoticons.

```
:D/NFP
 =-RRB-/NFP
^_^/NFP
```

3.3.2 Text Decoration

Punctuation serving as text decoration is also tagged NFP.

```
**/NFP update **/NFP
ON BOARD */NFP BOTH */NFP
 **********/NFP  Internet Email Confidentiality Footer   **********/NFP
```

3.3.3 Other Punctuation

Other punctuation not serving a canonical purpose if tagged NFP as well. In the first example it is marking a person signing off, and in the last it is indicating the hierarchy of items in a drop-down menu.

```
~/NFP Jason
@/NFP W.a.b.b.y
Click colors  >>>/NFP levels
```

3.4 Breed Names

Breed names are done as NN(S) rather than NNP(S) since capitalization is inconsistent across breed names in formal English to start, and even much more so in webtext.

```
yorkie/NN
dachshund/NN
Norwegian/JJ Forest/NN Cat/NN
Silkies/NNS
```

3.5 Separated Suffixes

Separated suffixes are tagged AFX.

```
male(s) -> male ( s/AFX )
```

3.6 Intensifiers

The intensifiers, such as 'fucking' are tagged according to position: JJ prenominally and RB otherwise.

```
I hate those fucking/JJ assholes
Those assholes fucking/RB did it again
```

3.7 Quoted Words

Quoted words behave syntactically as nouns, and so are tagged as such.

```
insert " a/NN " before the word " change/NN " and after the word " in/NN "
delete the " s/NN " from the word consolidation
```

3.8 Puns

Tokens should be tagged as the part of speech they are as long as it does not interfere with Treebank annotation, as in the following example describing a Vietnamese noodle soup shop.

```
Pho/NN -/HYPH nomenal/JJ !
```

3.9 JJ or VBG/VBN

The following have been standardized to VBG/VBN.

```
mind blowing/VBG
highly recommended/VBN
disappointed/VBN to
greatly appreciated/VBN
get married/VBN
```

The following has been standardized to JJ.

```
is concerned/JJ
```

3.10 Other Decisions

The following is a grab bag of POS decisions that have come up recently.

```
I <3/VBP Max 's
thou art/VBP
It was meh/JJ
RIP/UH
```

4 Treebank

4.1 Treebanking Typos, Misspellings, etc.

As with the other levels of annotation, sentences are bracketed with the correct structure regardless of incorrect spelling.

```
(S (NP-SBJ-1 they're feet)
   (VP  need
       (S (NP-SBJ-1 *)
```

```
            (VP  to
                (VP  be
                    (VP  done
                        (NP-1 *)))))

(S (NP-SBJ thei)
   (VP  r
       (ADJP-PRD energetic)
       (PP-LOC  at
               (NP night))))

(S (NP-SBJ the)
   (VP  re
       (PP-PRD  from
               (NP iraland))))
```

4.2 Initialisms, etc.

Are tagged as the function of the expanded phrase. Initialisms that resolve to an inherently adverbial phrase (ADVP, PP, XP-ADV) are annotated as ADVP, as 'atm' (for 'at the moment') is below.

```
(S (NP-SBJ-9 I)
   (VP  do n't
       (ADVP particularly)
       (VP  want
           (S (NP-SBJ-9 *PRO*)
              (VP  to
                  (VP  stay
                      (PP  with
                          (NP them))
                      (ADVP-TMP atm)
                      (ADVP either)))))))
```

Initialisms that are not performing a syntactic function are done as INTJ.

```
(INTJ roflmao)
```

'idk' often takes an argument. To preserve this argument structure, it is annotated as a the head of a verb phrase taking an empty subject: .

```
(S (NP-SBJ *PRO*)
```

```
(VP  idk
    (SBAR (WHADVP-1 where)
          (S (NP-SBJ I)
             (VP  can
                 (VP  go
                     (ADVP-DIR-1 T*)))))))
```

4.3 Pro-drop

The informal language in webtext tends to be more heavily pro-drop than formal English. Pro-dropped sentences are annotated as normal sentences with an empty subject.

```
(S (NP-SBJ *PRO*)
   (VP  Asked
       (PP-CLR  for
               (NP (NML bar  -  b - que)
                   sauce))))
```

In rare instances, this analysis causes empty categories to be adjoined, which is otherwise avoided.

```
(S (NP-SBJ (NP *PRO*)
           (SBAR-1 *EXP*))
   (VP  Didn't hurt
       (SBAR-1 that...)))
```

In the absence of other evidence, "got" is analyzed as a participle, and so gets treated as FRAG. Empty subjects are not inserted when both subject and auxillary are missing.

```
(FRAG (VP got (NP it)))
```

4.4 ADVP

The following was standardized to ADVP (note that this list is not exhaustive).

```
came up
```

4.5 FRAG

4.5.1 Contact Blocks

Blocks containing contact information are joined together by a FRAG node.

```
(FRAG (NP Rosario Gonzales)
      (NP Administrative Assistant)
      (NP ESSG)
      (NP (NP CCA-15)
           ,
          (NP 150301)
           ,
          (NP 153B09))
      (NP 281-514-3183)
      (NP rosario.gonzales@compaq.com)
       <
      (NP mailto:rosario.gonzales@compaq.com)
      (NP < mailto:rosario.gonzales@compaq.com >)
       >)
```

Similarly, blocks containing time stamps, e-mail sender names, etc., are annotated as FRAG as well.

```
(FRAG  |--------+-----------------------> | |
      (NP Tana.Jones@en | | | ron.com)
       | | | | | |
      (NP (NP 11/29/2000)
           | | |
          (NP-TMP 04:34 PM))
       | | | | |--------+----------------------->)
```

4.5.2 Timelines

FRAG is used for timelines as well.

```
(FRAG (NP 1555) :
      (S (NP-SBJ-1 Castling)
         (VP  is
             (VP  introduced
                 (NP-1 *))))
      .)
```

4.5.3 Announcements

Announcements are also annotated as FRAG.

```
(FRAG (NP WHO)
       :
      (NP Enron))

(FRAG (NP WHAT)
       :
      (NP (NP Happy Hour)
          (PP  for
              (NP John Suarez))))

(FRAG (ADVP WHEN)
       :
      (NP (NP Today)
          (PP-TMP  at
                  (NP 5 pm))))
```

4.5.4 Non-Native English

FRAG is also used for non-native English missing major grammatical elements, as in the sentence below (preceding sentence is included for context).

```
(S Hi I 'm from Brazil and I want to know of book 06 .)
(FRAG Is good or bad ?)
```

4.5.5 Other Constructions

"Here is to..."

```
(FRAG (ADVP here) (VP is (PP-PRD to getting back on track after Thanksgiving)))
```

4.6 INTJ

The following patterns are annotated as INTJ.

```
(INTJ Hi there)

(INTJ (INTJ hi)  (NP-VOC John))
```

4.7 PRT

The following collocations were standardized to PRT (note that this list is not exhaustive).

call in
going on
put it in
put down (insult)
put up with

4.8 QP

'Tens/hundreds of thousands of' has been standardized in this data to the following.

```
(NP (NP (QP tens of thousands))
    (PP  of
        (NP ...)))
```

Other QPs:

```
(NP (QP a couple) times)
(NP (QP nearly two) years)
(NP (QP 5 - 10) minutes)
```

4.9 WHNP

Nominal premodifiers in WHNP are annotated with NML; since in other cases wh-words are dominated by a WH node of some flavor, this in theory should be *WHNML. But this node label does not officially exist.

```
(WHNP (NML what size) horse)
(WHNP (NML what caliber) rifle)
```

'How X is too X?' is annotated with a WHNP to allow it to legally trace to subject position.

```
(SBARQ (WHNP-1 how rough) (SQ (NP-SBJ-1 *T*) (VP is (ADJP-PRD too rough)))?)
```

'What all X' has a WHADJP around 'what all'.

```
(SBARQ (WHNP (WHADJP What all) places) (SQ have you visited)?)
```

4.10 X

4.10.1 Non-linguistic Text

X is used to mark sentence units containing no linguistic information.

```
(X >------------------------------  |   | >------------------------------|)
(X ---->===}*{===<----)
```

4.10.2 Mistyped Text

Text entered more than once or in the wrong place is annotated with X.

```
(NP (NP Huge ammount)
    (PP  of
        (X time)
        (NP (NP wasted time)
             and
            (NP elevated blood pressure)))
    .)

(NP each and every  (X one)  one)

(SBARQ what (X the) s difference between)
```

4.10.3 Foreign Language Text

Text spans containing foreign language text are annotated with X.

```
(X A la guerre c'est comme a la guerre !)

(X Kitna soyega uthja yaar)
```

4.11 Dash Tags

4.11.1 CLF

Cleft sentences with 'that' annotated the same way as it-clefts.

```
(S-CLF (ADVP so)
       (NP-SBJ that)
```

```
(VP  was
    (ADVP-TMP-PRD (NP 2 days)
                  ago)
    (SBAR (WHADVP-1 that)
          (S (NP-SBJ I)
             (VP (VP called)
                 and
                (VP  left
                    (NP a message))
                (ADVP-TMP-1 *T*))))))
```

Cleft sentences with reduced rather than infinitival relatives have full relative structure added.

```
(SBAR  Since
      (S-CLF (NP-SBJ it)
             (VP  will
                 (VP  be
                     (NP-PRD Janet)
                     (SBAR (WHNP-1 0)
                           (S (NP-SBJ-1 *T*)
                              (VP  signing
                                  (NP the papers))))))))
```

4.11.2 CLR

The following collocations were standardized to use the CLR tag in this corpus (note that this list is not exhaustive).

choose from
come across
do with
get along with
get in touch with
go through
hear from
help with
listen to
look into
mean a lot to
order from

pay for
show for
use for
work for
work from
work with

4.11.3 DIR

The following collocations were standardized to use the DIR tag in this corpus (note that this list is not exhaustive).

come back
come here
come home
come in
come out of
get home
get home to
get on
get out
go anywhere
go away
go back
go down
go here
go home
go in
go out
go there
go to
head out
move here
move in
ran away
return home
walk around
walk away
walk in

4.11.4 ETC

NP-ETC is used in cases such as the following.

```
(S (NP-SBJ WE)
   (VP (VP DRIVE
            (ADVP AROUND)
            (PP-LOC IN
                    (NP MY VAN)))
        AND
        (VP SOLVE
            (NP MYSTERYS))
        AND
        (NP-ETC SHIT)))
```

4.11.5 LOC

The following collocations were standardized to use the LOC tag in this corpus (note that this list is not exhaustive).

be in the emergency room
dine in
see below

4.11.6 MNR

The following collocations were standardized to use the MNR tag in this corpus (note that this list is not exhaustive).

live in harmony
live in poverty
work hard

4.11.7 PRD

The following collocations were standardized to use the PRD tag in this corpus (note that this list is not exhaustive).

feel the same way
go wrong
stay away
stay outside

4.11.8 PRP

The following collocations were standardized to use the PRP tag in this corpus (note that this list is not exhaustive).

come for
(to) die for
thank you for

4.11.9 TPC

'What do you mean X.'

```
(SBARQ What do you mean (S-TPC I'm a pervert))
```

4.12 Other Decisions

'Hands down', 'thumbs up', etc. are done as NP with adjoined ADVP.

```
(NP-ADV (NP hands) (ADVP down))

(NP (NP two thumbs) (ADVP up))
```

Quasi-extraposition is done with S-ADV:

```
(S (NP-SBJ The Bush administration)
   (VP  would
       (VP  be
           (ADJP-PRD insane)
           (S-ADV to attack Iran))))
```

The title "Love's Labour's Lost" is analyzed as a passive.

```
(S-TTL (NP-SBJ-1 (NP Love 's)
                 Labour)
```

```
        (VP  's
            (VP  Lost
                (NP-1 *))))
```

'Make sure' and 'make clear' are done as follows.

```
(S (NP-SBJ He)
   (VP made
       (S (NP-SBJ *PRO*)
          (ADJP-PRD  sure
                     (SBAR that...)))))

(S (NP-SBJ He)
   (VP made
       (S (NP-SBJ (NP it)
                  (NP-1 *EXP))
          (ADJP-PRD clear)
          (SBAR-1 that...))))

(S (NP-SBJ He)
   (VP made
       (S (ADJP-PRD clear)
          (SBAR-SBJ that...))))
```

'A little more' heading an NP has an ADJP inserted to allow an NP premodifier.

```
(NP (ADJP (NP a little) more))
```

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.

[2] Justin Mott, Ann Bies, Colin Warner, and Ann Taylor. *Supplementary Guidelines for ETTB 2.0.* University of Pennsylvania, 2009.

[3] Ann Taylor. *Bracketing Switchboard: An Addendum to the Treebank II Guidelines*, 1996.

[4] 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.