SMS/Chat Treebank Guidelines

version1.3, March 31, 2013

Linguistic Data Consortium

Justin Mott, John Laury, Ann Bies and Colin Warner

Tokenization

The policy toward tokenization in informal SMS/chat data is to be editorial. That is, to reconstruct the author's intent in order to allow for coherent POS and Treebank annotation. This requires manual correction of the output of automated tokenization.

Contractions and possessives written without apostrophes are split into separate tokens. Likewise, novel informal contractions are split. Misspellings containing apostrophes are kept as single tokens to reflect their use.

```
Annes dog -> Anne s dog
Gregs ill -> Greg s ill
sup wit chu -> s up wit chu
whaddup? -> whad d up ?
Eek, I forgota writecha back -> Eek, I forgot a write cha back
I'ma swing by real soon -> I 'm a swing by real soon
How are you're analyses -> How are you're analyses  [no change]
```

Separate words accidentally typed with no space are split into separate tokens.

```
Lolhopefully -> Lol hopefully
Jerome,id. -> Jerome, id.
afterall -> after all
you got [for: go to] happy hour a lot -> you go t happy hour a
lot
```

Punctuation inside words is not tokenized separately.

```
Fll.a wants to know
```

Abbreviations containing slashes are kept as single tokens when they map to separate words.

```
And unfortunately ca n't go Sunday b/c I work all day
- No big projects w/o consulting
If you need to turn the a/c on , just turn the dial next to the
fridge above the couch
```

When typos involve the placement of a character in place of a space, that character is split off as its own token.

```
Theyrenplaying ac dc -> They re n playing ac dc
```

When typos or autocorrections cause spaces to appear in erratic locations, the erroneous space must be left intact. And characters incorrectly joined to another word will be split into a separate token.

```
I have sonic hi [for: something I] want to talk to you about ->
I have sonic h i want to talk to you about
```

POS

Emoticons get their own POS tag: EMO. Emoji characters are tagged as EMO as well. Note that processing restrictions require that emoji be "flattened" to an asterisk.

```
(:/EMO
://EMO
-__-/EMO
*/EMO (from: crescent moon emoji)
```

Semi-conventional web spellings and number/symbol substitutions are POS-tagged and annotated as though they are spelled normally.

```
tho/RB
abt/IN
thru/IN
b4/IN
#/NN
b/VB
ur/PRP$
u/PRP r/VBP
```

Initialisms are not split into their constituent parts. When used as exclamations in typed discourse, they are POS-tagged UH if the term stands alone or is an abbreviation for a complete sentence.

```
lol/UH
wtf/UH
ikr/UH
smh/UH
```

Otherwise, they are POS-tagged as they function within the sentence. Initialisms for PPs are tagged as RB.

```
btw/RB
imo/RB
```

A handful of intialisms are tagged as UH when occurring on their own but are tagged by function when occurring in a syntactic relationship with other constituents

```
Idk/UH !
vs. Idk/VBP what to say

ROFL/UH !
vs. rofl/VBG at those guys
```

Novel contractions are annotated by function.

```
s/VBZ up/RB wit/IN chu/PRP
whad/WP d/VBZ up/RB ?
```

Words are POS-tagged and annotated as though they are spelled correctly, as per the annotator's best understanding of the author's intent.

```
They seemed so ha[[y/JJ [for: happy]
but it always some/VBZ [for: comes] from so far away
I know texting I'd/VBZ [for: is] cheap and calling is better
we would go t/IN [for: to] the distributot
Everything still going ok there/RB '/. [for: ?]
```

For words that are accidentally typed with an internal space, the first (and any intermediate) token is POS-tagged GW and the last token is POS-tagged with the appropriate tag for the whole word.

```
:/GW )/EMO
```

When typos involve the placement of a character in place of a space, that character is POS-tagged XX.

```
They re n/XX playing ac dc.
```

When typos or autocorrections cause spaces to appear in erratic locations, characters incorrectly joined to another word will be POS-tagged as XX.

```
I have sonic/NN h/XX i/PRP want to talk to you about
```

Other material judged to be uninterpretable by the annotator (either occurring on its own or in isolation) will be POS-tagged as XX.

```
It is around d/XX the bend on the right,
Ap/XX
Ing/XX
```

Self-editing

Some authors realize their own typos or syntactic errors and retype a portion of their previous entry to correct themselves. Often the correction is preceded or followed by a *, which is POS-tagged /NFP. Multiple words are tagged as the context dictates. Single words are assumed to be mentions of words and are tagged NN.

```
yeah there certainly lots of little ones
there/EX were/VBD */NFP

they actually fo/VBP
do/NN */NFP
```

Treebank

Top-level node labels

In situations where sentence-final punctuation is present, the two separate sentence elements do not share any overarching node label. (Please note that in the following examples in many cases only partial trees are given to demonstrate the structure under discussion.)

```
(NP happy birthday!) (S Hope you're enjoying a day off!!)
```

It is common for authors to leave out sentence-final punctuation. Any complete sentences that do not have sentence-final punctuation between them will share an overarching node label.

```
(S (FRAG-SRED watching aj's game) (S looks a little wet).)
```

Pro-drop sentences

When only the subject is missing, S/SQ/SINV with *PRO* subject is used.

```
(S (INTJ yeah)
   (NP-SBJ *PRO*)
   (VP  should
       (VP  be
           (ADJP-PRD nice)
           (NP-TMP tomorrow))))
```

For empty subjects of relative clauses, infinitival clauses, etc., the default position is as the first daughter of S. Because there is no grammatical version of these classes with a surface subject, it is not possible to otherwise locate the subject's position with regard to other constituents. In pro-drop (and similar) sentences, the empty subject should be placed where it would occur in the non-pro-drop version of the sentence. That is, the following contrast in grammaticality implies the annotation below.

I just wanted to say hi. vs. *Just I wanted to say hi.

```
(S (NP-SBJ-1 *PRO*)
   (ADVP Just)
   (VP  wanted
       (S (NP-SBJ-1 *PRO*)
          (VP  to
              (VP  say
                  (INTJ-SEZ hi))))) .)
```

FRAG-SRED

This data set introduces a new node label: FRAG-SRED (for **S**entence **RED**uced), which borrows from the category S-RED from the BioMed guidelines (Warner et al., 2012). This node is used on sentence fragments missing copulas and/or auxiliary verbs.

With ADJP:

```
(FRAG-SRED (NP-SBJ *PRO*)
           (ADJP-PRD Awesome))
```

```
(FRAG-SRED (NP-SBJ *PRO*)
           (ADVP just)
           (ADJP-PRD great))

(FRAG-SRED (NP-SBJ *PRO*)
           (ADJP-PRD Sorry I missed your call))
```

With VP:

```
(FRAG-SRED (NP-SBJ *PRO*)
           (VP Waiting at doctor's office))

(FRAG-SRED (NP-SBJ You all)
           (ADVP-TMP still)
           (VP going))
```

Sentence units consisting of just an NP, PP or ADVP are not given a covering FRAG-SRED node:

```
(PP On my way :p)
(NP Sure thing)
(ADVP Here!)
```

The presence of additional constituents, as before, requires a covering FRAG node; these now get an -SRED tag, an empty subject and a -PRD tag:

```
(FRAG-SRED (NP-SBJ *PRO*) (PP-PRD On my way) (ADVP-TMP now) :p)
(FRAG-SRED (NP-SBJ *PRO*) (NP-PRD Sure thing) (NP-TMP this time))
(FRAG-SRED (NP-SBJ *PRO*) (ADVP-TMP still) (ADVP-LOC-PRD here) !)
```

Please note this change does **not** apply to small clauses, resultatives and other categories heretofore analyzed as an S missing an overt verb.

Coordination of FRAG-SRED

FRAG-SRED can be coordinated with either FRAG or S*. When coordinated with S* clauses, a S node dominates. Similarly, when coordinated with a (non-SRED) FRAG, there is a parent FRAG. Coordinated FRAG-SREDs are dominated by FRAG-SRED.

```
(S (S) and (FRAG-SRED))
(S (FRAG-SRED) and (SBARQ))
(FRAG (FRAG) and (FRAG-SRED))
(FRAG-SRED (FRAG-SRED) and (FRAG-SRED))
```

Interrogatives

In general, SQ should only be used to reflect that there has been syntactic movement typical of question formation (do-support, verb fronting). In the absence of those features (such as when auxiliaries are not present), we do not annotate with SQ. Instead FRAG-SRED is used.

```
(FRAG-SRED (NP-SBJ You)
           (ADVP-LOC-PRD there) ?)
```

```
(FRAG-SRED (NP-SBJ he)
           (PP-PRD in grad school) ?)

(FRAG-SRED (NP-SBJ you)
           (ADVP-TMP still)
           (ADVP-LOC-PRD home) ?)

(FRAG-SRED (NP-SBJ u)
           (ADJP-PRD alive)
           (ADVP-LOC out dere) ?)

(FRAG-SRED (NP-SBJ you guys)
           (PP-LOC-PRD on the road) ?)

(FRAG-SRED (NP-SBJ matt)
           (ADJP-PRD drunk) ??)

(FRAG-SRED (NP-SBJ *PRO*)
           (VP asking) ?)

(FRAG-SRED (NP-SBJ *PRO*)
           (VP reading anything good) ?)

(FRAG-SRED (NP-SBJ you)
           (ADJP-PRD sure) ?)
```

Wh-questions with missing material are annotated as SBARQ dominating FRAG-SRED.

```
(SBARQ (WHNP-1 what)
       (FRAG-SRED (NP-SBJ-1 *T*)
                  (ADVP-PRD up))  ?)

(SBARQ (WHADVP-1 Where)
       (FRAG-SRED (NP-SBJ u)
                  (PP-LOC-PRD  at
                               (ADVP *T*))) ?)
```

INTJ

INTJ nodes in previous data were considered "invisible" to the syntax of the surrounding structure across the board. As such, their presence never forced any additional structure. In SMS/chat data, however, interjections are used much more often and are usually "meaningful." To capture this, the presence of an INTJ node can force the use of a node (FRAG) to cover the INTJ and the other constituent.

Old Policy:

```
(PP (INTJ yeah) on campus)
(ADVP unfortunately (INTJ yes))
```

New policy:

```
(FRAG (INTJ yeah) (PP on campus))
(FRAG (ADVP unfortunately) (INTJ yes))
```

Interjections inside S* clauses do not project extra structure unless there is a constituent (such as a conjunction) indicating additional structure. This is done in part because it is difficult to distinguish filler vs. non-filler interjections and in part because this pattern is extremely common and would generate even more FRAG nodes.

```
(S (INTK Ok) see you then.)
(UCP (INTJ yes) and (S unfortunately he loves it))
```

SMS/chat data exhibits a large number of items being arguably used as interjections. In general, we are conservative in annotating things as INTJ and prefer to annotate according to lexical meaning when possible. Some items that are tagged as INTJ include the following.

```
Haha
LOL
xoxo
xx
Oh shoot
Atta boy
Damn
K
Shit!
```

"Right" and "okay" (and variants such as OK, ok, kay, K) can be annotated as either INTJ or ADJP. The default is INTJ when occurring on their own or outside of argument structure, but they can also be ADJP when acting as predicate adjectives or modifying nouns.

```
(FRAG But at like eight (INTJ right)?)
(S Silva has never been beat (INTJ right))
(S (INTJ okay), I thought I heard your voices chatting about
lunch.)
(INTJ (INTJ Haha), (INTJ okay))
(FRAG-SRED (INTJ ok), cool)

(S You're (ADJP-PRD right)!!!!)
(S It was (ADJP-PRD okay))
(SQ Is 7 (NP-PRD an ok time)?)
```

The following items are examples of potential INTJs that we annotate as their lexical value.

```
Lame
Awesome
awkward
fair
DUMB
Same!!
```

```
Really?
Kiss
Sweet
No problem
Wait
Look
Say
Word
```

Incomplete Sentence Units

Statements that cut off before being complete must be annotated as two separate sentence units. The annotator can use insight from one section to determine what structure is present in another, however.

```
(FRAG (INTJ no), but)

(FRAG (S I know texting I'd cheap and calling is better) but)

(S (NP-SBJ my day) (VP slipped (ADVP away))))
(PP from (NP me).)

(FRAG (S It's also really bleak and  violent) (ADVP so))
(S it might not be your cup of tea.)
```

In cases where a single word is split between two SUs, the incomplete segments are under an X-node. The X-nodes containing partial words are not treated as though they contain a meaningful word.

```
(S Please (VP (FRAG ca)))
(X lll)

(FRAG (SBAR If you are bored) (X a))
(FRAG (X nd) (VP want to send a letter to say hi) (S I'm sure they
would enjoy it).)
```

Use of X nodes

When typos involve the placement of a character in place of a space, that character is placed under an X-node.

```
(S (NP-SBJ They)
   (VP  re
       (X n)
       (VP  playing
           (NP ac dc))))
```

When typos or autocorrections cause spaces to appear in erratic locations, the characters incorrectly joined to another word are placed under an X-node.

```
(S  I have
   (NP (NP sonic)
```

```
(X h)
(SBAR (WHNP-1 0)
      (S i want to talk to you about
         (NP-1 *T*))))
```

In cases where a word is repeated, the first instance is placed under an X node.

```
(SQ Do you mostly play (X at) (PP-LOC at church) ?)
```

“Sentences” consisting only of an emoticon or other punctuation are annotated as X.

```
(X :-RRB- )
(X * ? )
```

Messaging metadata that crops up occasionally is annotated with X.

```
(NP the (X 2 / 2) other)
```

Other

Initialisms interacting with other constituents are tagged by function as much as possible. Note that 'idk' is treated as a pro-drop verb. Also note that splitting 'idk' into its constituent parts would not yield the negative particle.

```
(S (NP-SBJ *PRO*) (VP idk what to say)
(S I am (VP rofl at those guys)
```

The pattern *I’m a [verb]* is annotated as a raising construction.

```
(S (NP-SBJ-1 I)
   (VP  'm
       (S (NP-SBJ-1 *)
          (VP  a
              (VP  swing
                  (ADVP-DIR by)
                  (ADVP-TMP real soon)))))
```

References

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre. 1995. *Bracketing Guidelines for the Treebank II-style Penn Treebank Project.* University of Pennsylvania.

Justin Mott, Ann Bies, John Laury, Colin Warner. 2012. *Bracketing Webtext: An Addendum to the Penn Treebank II Style Bracketing Guidelines*. Linguistic Data Consortium, Philadelphia.

Justin Mott, Colin Warner, Ann Bies, Ann Taylor. 2009. *Supplementary Guidelines for ETTB 2.0*. Linguistic Data Consortium, Philadelphia.

Colin Warner, Ann Bies, Christine Brisson, Justin Mott. 2004. *Addendum to the Penn Treebank II Style Bracketing Guidelines: BioMedical Treebank Annotation*. Linguistic Data Consortium, Philadelphia.

Colin Warner, Arrick Lanfranchi, Tim O'Gorman, Amanda Howard, Kevin Gould, Michael Regan. 2012. *Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines*. Institute of Cognitive Science, University of Colorado at Boulder.