Bracketing Raw Switchboard Text

1 “Sentence” units

The E_S, N_S notation is used to indicate complete and incomplete “sentences” as given by the dysfluency annotation. These may be ignored at the bracketing stage at the discretion of the annotator. When this is done, the E_S or N_S is treated as punctuation and attached as high as possible. If a sentence has been wrongly divided across a turn at the dysfluency stage, it is not possible for the annotators to correct it, and thus it may remain as an error.

( (FRAG but
   (INTJ uh)
   ,
N_S
   (NP (NP (ADVP just about)
        any kind)
       (PP of
        (NP music))))
   .
E_S))

( (SQ (EDITED (SQ Is
       (NP-SBJ-UNF i-)))))
   ,
N_S
EDITED (RM []
   (SQ-UNF are)
   ,
   (IP +))
are
(RS ])
(NP-SBJ they)
(ADJP-PRD good)
?
E_S))

2 Individual words and phrases

• Fillers
  Fillers such as uh, um, etc. are bracketed (INTJ uh).

  ( (S (INTJ Um)
(INTJ well)

(NP-SBJ you)

(VP talked

(PP about

(INTJ uh)

(NP volcanos)))

E_S))

• Editing terms

Editing terms, extremely commonly I mean but sometimes also excuse me and a few other things, are labelled PRN. Within the PRN brackets these are bracketed as complete sentences. The usually present following comma is enclosed within the PRN.

( (S (PRN (S (NP-SBJ I)

(VP mean))

,)

(NP-SBJ they)

(VP do

(VP generate

(NP (NP a lot)

(PP of

(NP soot))))))

E_S))

• Discourse markers

Most discourse markers, like fillers, are labelled INTJ. This includes the extremely common well and like, but also occasional others like now, see, say, actually, anyway, etc.

( (S (INTJ Well)

,)

(NP-SBJ it)

(VP ’s

(ADJP-PRD hard

(SBAR (WHNP-1 0)

(S (NP-SBJ *))

(VP to

(VP say

(NP *T*-1))))))

).

E_S))
The discourse marker you know is labelled PRN like editing terms and is bracketed inside the PRN with the preceding and following commas (if present) enclosed.

```
( (FRAG and
   (PRN ,
    (S (NP-SBJ you)
     (VP know))
    ,)
   (NP spots)
   (PP-LOC (ADVP all)
     over
     (NP them))
   .
   E_S))
```

So can be either a discourse marker or a conjunction. As a discourse marker it is labelled INTJ, as a conjunction it is unlabelled.

```
( (SQ (INTJ So)
   .
   Have
   (NP-SBJ you)
   (VP heard
    (PP about
     (NP Saturns)))
   ?
   E_S))
```

```
( (S so
   (NP-SBJ we)
   (VP 'll
     (VP pick
      (PRT up)
      (NP that)))
   .
   E_S))
```

Combinations of INTJs like oh really, oh yeah, etc. are each labelled individually as INTJ and the highest label is also INTJ.

```
( (INTJ (INTJ Oh)
   ,
   (INTJ really)
   .
   .
   .)
```
E_S))

( (INTJ (INTJ oh)
    (INTJ yeah)
    ,
    E_S))

• Conjunctions

Conjunctions are treated in the traditional way; that is, they are left bare if they are a single
word, but labelled CONJP when multi-word. Following dysfluency practice, and then and
and so are treated as multi-word conjunctions.

( (S but
    (NP-SBJ I)
    (ADVP certainly)
    (VP work
        (ADVP-LOC near
            (NP a city))))
    ,
    E_S))

( (S (CONJP and so)
    ,
    (NP-SBJ the best bet)
    ,
    (EDITED (RM [])
        (EDITED (RM [])
            (ADVP-TMP-UNF any)
            ,
            (IP +))
            (ADVP-TMP-UNF anymo-)
            ,
            (RS [])
            (IP +))
        (NP-TMP these days)
        (RS [])
        (VP are
            (NP-PRD compact disk))
        ,
    E_S))

When an entire incomplete sentence is made up of a conjunction (plus perhaps fillers or other
junk) the highest label is S-UNF.

( (S-UNF and
(INTJ uh)
,  
N_S))

( (S-UNF (CONJP and then)
,  
N_S))

• Continuers and assessors of all types, uh-huh, huh, really, exactly, right, yeah, oh, okay, etc. are labelled INTJ and constitute a “sentence” on their own.

( (INTJ Yeah
,  
E_S))
( (INTJ okay
,  
E_S))
( (INTJ Uh-huh
.  
E_S))
( (INTJ Right
.  
E_S))

• Asides
  Asides which interrupt the flow of a sentence are fully bracketed inside PRN.

( (SBARQ (WHNP-1 what)
 (SQ ’s
  (NP-SBJ that
   (EDITED (RM [])
    (ADJP-UNF real)
    ,
    (IP +))
   (INTJ oh)
   ,
   (ADJP not real
    (RS []
     old)
   (PRN ,
    (S (NP-TPC my wife)
     (NP-SBJ she)
     (VP is

5
(ADJP-PRD old)
(ADVP too))
)
dancer)
(NP-PRD *T*-1)
,
(INTJ uh))
,
E_S))

3 Restarts

The linguistic theory behind restarts is that the speaker, on deciding that s/he wants to fix something already uttered, goes back one constituent and restarts from there. For this reason we try to make restarts sisters of the replaced constituent and have them contain a single constituent which is sometimes unfinished (and therefore labelled -UNF). There are two problems with implementing this approach, however: (1) exactly what counts as a "constituent" for restarting is not particularly clear; and (2) the way Treebank brackets things doesn’t always provide bracketing at the appropriate level to do this accurately. We have therefore had to make some compromises, while to the extent possible keeping the system both internally consistent and linguistically plausible.

Summary of restart policy

- restarts of phrases (any label that doesn’t start with S) are sisters
- restarts of clauses (any label that starts with S) are embedded
- if the restart starts at a point where in normal Treebank bracketing there is no label immediately to the left, there is no label inside the restart.
- any label at top level is treated like a clause and the restart is embedded.

3.1 The internal structure of restarts

The brackets indicating restarts in the dysfluency annotation appear in the syntactic bracketing as (RM |) (IP +) and (RS |) (although these may be ignored by the annotators if in their opinion they are in the wrong place). The part of the restart between (and including) (RM |) and (IP +) is labelled EDITED. It indicates the part of the restart to be ignored. (RS |) (except in restarts without repair, see below) is treated like punctuation and is attached at the highest possible level.

```xml
(S (EDITED (RM |)
   (NP-SBJ I)
   (IP +))
  (NP-SBJ I)
  (RS |))
```
(VP do
  n't
  (VP care)))))

The nesting of EDITEDs follows that of the original restart in left-branching restarts (the most common kind), while other types are more likely to be treated as sisters.

- left-branching restart

  [ [ I + I ] + I ] really don't care

  ( (S (EDITED (RM []
    (EDITED (RM []
      (NP-SBJ I)
      (IP +))
    (NP-SBJ I)
    (RS []
      (IP +))
    (NP-SBJ I)
    (RS []
      (VP do
        n't
        (VP care))))))

- center-embedded restart

  [ they really + [ they + they ] really ] don't care

  ( (S (EDITED (RM []
    (S-UNF (NP-SBJ they)
      (ADVP really))
    (IP +))
  (EDITED (RM []
    (NP-SBJ they)
    (IP +))
  (NP-SBJ they)
  (RS []
    (ADVP really)
    (RS []
      (VP do
        n't
        (VP care))))))

All parts of a restart without repair are included inside the EDITED brackets, that is, including the (RS []).
( (S (NP-SBJ I)
   (VP bought
      (NP a cat)
      (EDITED (RM [])
         (PP-UNF in)
         (IP +)
         (RS []))
      (NP-TMP yesterday)))))

The dash-tag -UNF is used to indicate constituents which are unfinished, (whether inside or outside of restarts). -UNF is attached to the lowest possible constituent that can be identified with confidence.

( (S and
   ,
   (INTJ uh)
   ,
   (NP-SBJ it)
   (VP ’s
      (EDITED (RM [])
         (EDITED (RM [])
            (PP-UNF in)
            ,
            (IP +))
         (PP in
            (NP-UNF the very)
            (RS [])
            (IP +)))))

   (PP-LOC-PRD in
      (NP the
         (RS []))
      mountains))

E_S))

( (S (INTJ Uh)
   ,
   (ADVP-TMP sometimes)
   (EDITED (RM [])
      (S (NP-SBJ it)
         (VP-UNF is))
      ,
      (IP +))
   (NP-SBJ there)
   (VP is
      (RS []))

8
(NP-PRD no alternative))
.
E_S))

((SBAR (INTJ uh)
 ,
 (ADVP especially)
 (EDITED (RM []
 (SBAR-ADV-UNF if)
 ,
 (IP +))
 if
 (RS []
 (S (NP-SBJ it)
  (VP 's
   (PP-TMP-PRD after
     (NP an acute illness))))))
 .
 E_S))

((S (CONJP and then)
 (NP-SBJ she)
 (VP decided
  (EDITED (RM []
   (SBAR-UNF that)
 ,
   (IP +))
 (INTJ uh)
 (PRN ,
  (S (NP-SBJ she)
   (VP 's
    (ADJP-PRD (NP (QP eighty five)
       years)
      old))))
 ),
 (SBAR that
 (RS []
 (S (ADVP-TMP eventually)
   ...
))))))

((S (EDITED (RM []
 (S-UNF and
   (NP-SBJ it))
 ,
 (IP +)))
and
(NP-SBJ it)
(RS ])
(VP makes
(S (NP-SBJ you)
  (VP do
    (NP it)))
).
E_S))

( (S (EDITED (RM []))
  (S (NP-SBJ I)
    (VP think
      (SBAR that
        (S-UNF (ADVP probably)))))
  ,
  (IP +))
(NP-SBJ I)
(VP think
  (RS ])
  (SBAR 0
    (S (NP-SBJ most companies)
      (ADVP-TMP now)
    ,
    (ADVP (ADVP as far)
      (PP as
        (NP entry
          ,
          (PRN (S (NP-SBJ you)
              (VP know))
           ,
           exams)))
      (VP require
        (NP that))))
  ).
E_S))

3.2 Restarts of constituents smaller than a clause (NP, PP, VP, ADJP, etc).

Restarts of non-clausal constituents are treated as sisters (except if the constituent is at top level, see below). If the restart starts at the beginning of a labelled constituent, the constituent inside the EDITED brackets has the same label (with -UNF if it is unfinished). If the part inside would not normally have a bracket to its left, as for instance, conjunctions or green in the big green dog, there is no labelled bracket inside the EDITED brackets either.
( (S (INTJ Well)
   ,
   (NP-SBJ I)
   (VP think
      (SBAR that
         (S (EDITED (RM [])
            (NP-SBJ it)
            ,
            (IP +))
         (NP-SBJ it)
         (RS [])
         (VP varies
            (EDITED (RM [])
               (PP-UNF on)
               (IP +))
            (PP on
               (RS [])
               (NP the individual basis))))))
   .
   E_S))

( (S but
   (NP-SBJ it)
   (VP does
      n't
      (VP stop
         (EDITED (RM [])
            (NP many people)
            (IP +))
         (NP-1 many people)
         (RS [])
         (EDITED (RM [])
            (PP-UNF from)
            ,
            (IP +))
         (INTJ uh)
         ,
         (PP from
            (RS [])
            (S-NOM (NP-SBJ *-1)
               (VP doing
                  (NP it))))))
   .
   E_S))
( (S (NP-SBJ I)
   (VP know
      (SBAR that
         (S (NP-SBJ certain states)
            (ADVP actually)
            (VP sell
               (EDITED (RM [])
                  (NP-UNF their)
               ,
               (IP +))
               (INTJ uh)
               ,
               (NP their
                  (RS [])
                  lists)))))))))

( (S (NP-SBJ Those aerobic classes)
   (EDITED (RM [])
      (VP-UNF are)
      ,
      (IP +))
   (VP are
      (RS [])
      (ADJP-PRD tough)
      (EDITED (RM [])
         (NP-TPC-UNF the)
      ,
      (IP +))
   (NP-TPC the
      (RS [])
      high impact))
   .
   E_S))

( (S (NP-SBJ-1 I)
   (VP 'll
      (VP be
         (ADJP-PRD able
            (S (NP-SBJ *-1)
               (EDITED (RM [])
                  (VP-UNF to)
               ,
               (IP +))
            (VP to
               (RS [])
               .
               E_S))

12
(VP do
  (NP it))))))

(S (NP-SBJ I)
  (VP ’d
    (VP be
      (EDITED (RM []
        (ADJP-PRD afraid)
      ,
      (IP +))
    ,
    (ADJP-PRD afraid)
    (RS [])))))

(SBARQ (EDITED (RM []
  (WHNP Who)
  ,
  (IP +))
  (WHADVP-1 how)
  (RS [])
  (SQ do
    (NP-SBJ you)
    (VP know
      (NP that)
      (ADVP-MNR *T*-1)))
  ?
  E_S))

(S and
  (NP-SBJ it)
  (VP depends
    (PP on
      (SBAR (EDITED (RM []
        (WHADJP-UNF how)
      ,
      (IP +))
      (WHADJP-1 how
        (RS [])
        sick)
      (S (NP-SBJ the person)
        (VP is
          (ADJP-PRD *T*-1))))
      (ADVP too))
    .))

13
Single word conjunctions have no constituent label so are not labelled inside restarts.

```
( (S (EDITED (RM []))
   (EDITED (RM [])
      and
      (IP +))
   and
   ,
   (RS []))
   (IP +))

and
,
(RS [])
(INTJ uh)
,
(EDITED (RM []))
   (S (NP-SBJ I)
      (VP-UNF ?m))
   ,
   (IP +))
(NP-SBJ I)
(VP 'm
   (RS [])
   (ADJP-PRD really gung ho
      (PP for
         (NP it)))))
.
E_S))
```

In the following examples, as well, there is no labelled bracket inside the restart because the restart does not begin at the beginning of constituent (as defined by Treebank).

```
( (S So
 (NP-SBJ I)
 (VP think
   (SBAR 0
    (S (NP-LOC (NP some
       (EDITED (RM [])
          place
         ,
         (IP +))
         (PRN (S (NP-SBJ you)
            (VP know)))
      ,)
   places)
```
(RS ])
(PP like
(NP that)))
...)))))))

(S (EDITED (RM []))
(S (NP-SBJ I)
(VP-UNF 'm))
,
(IP +))
(NP-SBJ I)
(RS ])
(VP guess
(SBAR 0
(S (NP-SBJ I)
(ADVP probably)
(VP put
(PRT on)
(NP (QP about
(EDITED (RM []))
ten
,
(IP +))
nine or ten)
(RS ])
pounds)))))))

(S (EDITED (RM []))
(S-UNF but
(NP-SBJ I))
,
(IP +))
(PRN (S (NP-SBJ you)
(VP know))
,
but
(NP-SBJ my kids)
(RS ])
(VP are
(ADVP only)
(NP-PRD elementary
(EDITED (RM []))
grades
,
(IP +))
levels
    (RS ])
  (ADVP-TMP right now))
E_S))

Non-clausal restarts at top level are embedded just like clauses (see below).

  ( (NP (EDITED (RM []))
      (EDITED (RM []))
        (NP-UNF A)
        (IP +))
    (NP-UNF a
      (RS ])
    good)

      (IP +))

    a good
    (RS ])
southern solution

    .
E_S))

  ( (INTJ (EDITED (RM []))
      (INTJ-UNF Ye-)
      ,
    (IP +))

    yeah

    .
    (RS ])
E_S))

3.3 Restarts of clauses

Ideally repairs of clauses would be treated as sisters of the reparandum just as with phrase level restarts (see above). However, this is not possible because of the way Treebank brackets clauses. Specifically, in both top level and embedded clauses, there can be material inside the clause (as defined by Treebank) preceding the restart making it impossible to properly make them sisters. Therefore all clause level restarts are embedded.

  ( (S (INTJ Uh)
      ,
    (ADVP-TMP sometimes)
    (EDITED (RM []))
    (S (NP-SBJ it)
      (VP-UNF is))


16
(IP +)
(NP-SBJ there)
(VP is
  (RS ])
  (NP-PRD no alternative))
.
E_S))

((S But
  (EDITED (RM []
    (S (NP-SBJ I)
      (VP-UNF 've
        ,
        (ADVP just)
        (NP-ADV myself)))
    (IP +))
  (NP-SBJ I)
  (ADVP just)
  (VP notice
    (S (NP-SBJ myself)
      (RS ])
      (ADVP just)
      (EDITED (RM []
        (ADJP-UNF really)
        ,
        (IP +))
      (ADJP-PRD real
        (RS ])
        soft)))))
E_S))

((S (NP-SBJ she)
  (VP said
    (SBAR that
      (S (EDITED (RM []
        (S (NP-SBJ all men)
          (VP-UNF will))
        ,
        (IP +))
        (NP-SBJ all people)
        (RS ])
        (VP will
          (VP be
            (ADJP-PRD free))))))))))
( (S (NP-SBJ she)
  (VP said
    (S (NP-TMP one day)
      (EDITED (RM [])
        (S (NP-SBJ all men)
          (VP-UNF will))
        ,
        (IP +))
      (NP-SBJ all people)
      (VP will
        (VP be
          (ADJP-PRD free))))))

( (SBARQ (INTJ Now)
  (EDITED (RM [])
    (SBARQ (WHNP what)
      (SQ-UNF 's))
    ,
    (IP +))
  (WHNP-1 what)
  (SQ 's
    (RS []
      (NP-SBJ toning)
      (NP-PRD *T*-1))
  ?
  E_S))

( (S (NP-SBJ she)
  (VP says
    (SBAR (SBAR-ADV if
      (S (NP-SBJ she)
        (VP fails
          (NP the exam)))))
    (EDITED (RM [])
      (SBAR-UNF that)
    ,
    (IP +))
    that
    (S (NP-SBJ she)
      (VP will
        (VP jump
          (PRT off)
          (NP a bridge))))))
( (S (NP-SBJ 1)
   (VP think
     (SBAR (SBAR (EDITED (RM []))
       (SBAR-UNF that)
     ,
     (IP +))
   that
   (RS ])
   (S (SBAR-ADV (EDITED (RM []))
     (SBAR-ADV-UNF if)
   ,
   (IP +))
   if
   (RS ])
   (S (NP-SBJ there)
     (VP is
       (NP-PRD (NP a problem)
         (PP with
           (NP the school system)))))
..))))))

( (...
or
   (SBAR-PRP (EDITED (RM []))
     (SBAR-PRP-UNF because)
   ,
   (IP +))
because
   (RS ])
   (S (NP-SBJ-2 you)
     (VP wo
     n’t
     (VP be
       (ADJP-PRD able
         (S (NP-SBJ -*2)
           (VP to
             (VP *?*)))))))
.
E_S))

( (S (NP-SBJ-1 you)
   (RS ])
   (VP want

19
(S (NP-SBJ *-1)
   (VP to
      (VP think
         (SBAR (EDITED (RM [])
            (SBAR-UNF that)
             ,
             (IP +))
          that
          (RS []))
         (S (NP-SBJ you)
            (VP ’re
               (VP working
                ...))))))

( (S (INTJ Oh)
   ,
   (S-ADV (EDITED (RM [])
               (SBAR-ADV if
                  (S (NP-SBJ you)
                    (VP-UNF do)))
                   ,
                   (IP +))
                 if
                (S (NP-SBJ you)
                  (VP do
                   (RS []))
                  (NP aerobics)))))
...
)

( (SBARQ (INTJ Now)
   (EDITED (RM [])
    (SBARQ (WHNP what)
     (SQ-UNF ’s))
    ,
    (IP +))
   (WHNP-1 what)
   (SQ ’s
    (RS []))
   (NP-SBJ toning)
   (NP-PRD *T*-1))
? E_S))

( (SBAR-PRP because
   (S (NP-SBJ you)
(VP 're
 (VP endangering
  (NP (NP everybody 's)
   lives)
 (SBAR-ADV (EDITED (RM []))
  (SBAR-ADV if
   (S-UNF (NP-SBJ you)))
  (IP +))
 if
 (S (NP-SBJ it)
  (VP 's
   (RS []))
 (NP-PRD (NP something)
   (PP like
    (NP that)))))))
.
E_S))