Addendum to the POS Tagging Guidelines

1 General changes

- The tag **/TO** is restricted to infinitival *to*. When *to* is used as a preposition it will be tagged **/IN** like all other prepositions.

  ```
  I want to/TO go.

  I went to/IN the store.
  ```

- The contracted verb form *'s* can stand for either *is* or *has*. To eliminate this ambiguity we are introducing two new tags: **/BES** for *'s* when it is contracted from *is* and **/HVS** when it is contracted from *has*.

  Thus *'s* can now be tagged in four different ways:

  ```
  He 's/BES a big boy now.

  She's/HVS got lots of money.

  That is Lee 's/POS bike.

  Let 's/PRP go home now.
  ```

2 Tagging Switchboard

2.1 The /XX tag

The **/XX** tag is used for partial words, but only when you cannot figure out from the context what the word is. If it is clear what the word is, tag it as usual.

```
[ we/PRP ]
got/VBD
[ one/CD ]
[ -s/XX ]
,/,
[ one/CD cut/VBN out/RP ]
on/IN
[ the/DT table/NN ]
saw/NN ,/,
```

When you know or can figure out what a word is but there isn't enough context to allow you to decide what tag to give it (in the case of words with multiple possibilities), use all the possibilities separated by a vertical slash.

2.2 The /UH tag

In tagging switchboard material, you will need to make heavy use of the tag **/UH** "interjection, which isn't used much when tagging written English. The following classes of words are tagged **/UH**.

- Fillers: *uh-huh, uh, um, huh, o, oh, ooh,* etc.
- Exclamations: *wow, boy* (as in *boy, i really like pizza*), *god, goodness, gosh, hey, jeez, my* (as in *my, you do have big teeth, grandmother*), *bam, bang,* etc.
- Yes and no answers: *yeah, yes, no, nope, nah,* etc.

  No as in *no longer* remains **/RB**
- Greetings and adieus: *bye-bye, bye, hello, hi,* etc.

  Good morning, etc. are tagged according to the class of the indvidual words, not as **/UH**: *Good/JJ morning/NN.*
- Continuers and assessors: *exactly, really* (*oh really* is *oh/UH really/UH*), *right, yeah, sure,* etc.

  This category includes generally one-word turns (except *oh really*) that serve only to reassure the other speaker that you're listening and basically agree. Don't tag these words UH if they are used in a sentence with their usual meaning (*this is exactly/RB (not exactly/UH) what I want*)
- Discourse markers: *well, actually, see* (as in *See, I told you I was right*), *sure, like* (as in *He's, like, a real cool dude*), *so* (as in *So, whadya wanna do?*), *etc.*

  To get the idea about this category, read the section on discourse markers in the Dysfluency Stylebook. All discourse markers EXCEPT *you know* will be tagged as **/UH** (*you know* is *you/PRP know/VBP*). Again the **/UH** tag is not used when the words are used in their normal meaning (*I don't actually/RB know the answer, He paints well/RB*). It is generally quite easy to recognize discourse *well*, but *actually* can be a bit difficult. Often it is utterance initial (or close to) and followed by a pause (i.e. a comma). There are examples in the Dysflency Stylebook. *So* is also sometimes a discourse marker. Again details are in the Stylebook.

2.3 Repeated words

The Switchboard Corpus is full of restarts, which means lots of repeated words. All identical words in a restart should be tagged the same. The case where this is an issue will only arise when the tag for a word is affected by following context. Thus in the example below, both the first and the second *what* should be tagged **/WDT** since the second *what* makes it clear that this is the proper reading.

```
[ What/WDT ]
,/,
[ what/WDT type/NN ]
of/IN
[ art/NN ]
do/VBP
[ you/PRP ]
focus/VB on/IN ?/.
```

2.4 The GW tag

GW is for correcting typos which affect tokenization of words; that is, when the transcriber spelled as two separate words something that should be spelled as one word. There are some very clear cases, like *a like* for *alike*, *back ground* for *background*, and many less clear cases. The final arbitrator for whether two items go together to make one word will be the electronic dictionary, Noah (American Heritage Electronic Dictionary Copyright 1991 by Houghton Mifflin Company). If you look the collocation up in Noah and it gives it as one entry spelled with only hyphens between the parts then you can use GW. So if you look *well rounded* up in Noah it gives the following:

```
[well-round-ed]
 (we_l'roun'di_d)

(ADJECTIVE).
    1. Comprehensively developed: ``a well-rounded scholar.''
    2. Having a shapely figure.
```

Since the parts are separated only by hyphens, this is a possible place to use GW. If Noah does not reckonize the collocation, or recognizes it but spells it with spaces rather than hyphens between the parts, this indicates that GW is NOT possible. Thus, while *well rounded* appears in Noah as *well-rounded* and can therefore be tagged *well/GW rounded/JJ*, *well dressed* is not recognized, so must be tagged *well/RB dressed/JJ*. It is not, however, necessary to tag hyphenated collocations with GW, so *well rounded* can also be tagged *well/RB rounded/JJ*. In general, try to use GW only when it is not possible to acurrately tag the two parts separately.

2.4.1 Places where GW should be used

- separated prefixes
  - *pre-*

    pre/GW med/JJ student/NN
    pre/GW regime/NN crimes/NNS
  - *mini-*

    mini/GW series/NNS
    mini/GW skirt/NN
    But note that when standing alone as a short form for miniskirt (she wore a mini/NN), it is a noun
  - *inter-*

    inter/GW state/NN
    inter/GW United/NNP States/NNPS
  - *ex-*

    ex/GW husband/NN
    ex/GW boy/NN friend/NN

    But note that *my ex* would be *my/PRP$ ex/NN*
  - *semi-*

the semi/GW final/JJ game/NN
a semi/GW nice/JJ dinner/NN

But note that *semi*, like *mini*, can also be used as a noun, i.e., a type of truck (*he drives a semi/NN*)

- *non-*

  non/GW fiction/NN books/NNS

- Nouns derived from verb-particle collocations are considered single words and thus when separated should be joined with GW

  the kids are drop/GW outs/NNS
  he's a real cut/GW up/NN
  we had a break/GW down/NN
  all kinds of nice write/GW ups/NNS about it
  did you do your work/GW out/NN this morning?
  a sixty/CD five/CD percent/NN turn/GW out/NN
  the reason for the split/GW up/NN

- Two word verbs should be joined by GW.

  Do you tent/GW camp/VB or do you have a camper
  If your child back/GW talks/VBP
  I substitute/GW teach/VBP
  I always over/GW pay my deductions
  You can easily over/GW do/VB it
  He always dry/GW cleans/VBP his shirts
  (BUT NOTE: dry/JJ cleaner/NN, dry/JJ cleaning/NN)
  They drug/GW test/VBP and it's not random
  (BUT NOTE: they gave him a drug/NN test/NN; they do drug/NN testing/NN)
  I was trying to saltwater/GW fish/VB
  (BUT NOTE: I went surf/NN fishing/NN
  The dog was being house/GW sit/VBN
  They were down/GW grading/VBG all these other ones

2.4.2 Places NOT to use GW and what to do instead

- Do not use GW for things like *R V -ers*. Label each part of the collocation according to what it would be if it were written out in full and label the inflection (the part after -) for the whole collocation.

  R/NN V/NN -er/NN
  R/NN V/NN -ers/NNS
  R/NN V/NN -ing/VBG
  M/NNP and/CC M/NNP -s/NNPS

  but T/GW V/NN

  Don't use GW for things like *F B I* or *V C R*, rather each letter is tagged as it would be if the word was written out. If when written out the letters would be capitalized, then label them NNP, but if not, then use NN.

V/NN C/NN R/NN
A/NN T/NN M/NN
P/NN C/NN
F/NNP B/NNP I/NNP
U/NNP S/NNP of/IN A/NNP
D/NNP C/NNP

Do not use GW to join anything up to a *'s* possessive marker. These are supposed to be separate tokens.

the F/NNP B/NNP I/NNP 's/POS star investigator
John/NNP 's/POS kitten/NN

- In general it won't be correct to join together prepositions with GW. Although *into* and *onto* are indeed different from *in to* and *on to*, in most cases both will be possible in the same contexts. The difference between *he put it into the bag* and *he put it in to the bag* is only one of stress, which we, of course, don't have access to. So unless it's absolutely impossible to read the sentence with a stress on each preposition, do not join them.

I went in/IN to/IN a hardware store
what would you be most interested in/IN getting in/IN to/IN
are you rich or poor or in/IN between/IN
he moved on/IN to/IN L A
the ability to move on/IN to/IN some new technology

- collocations with the "suffix" *wise* should be tagged separately with *wise* tagged as an adverb.

you probably pay more percentage/NN wise/RB
he's doing well popularity/NN wise/RB
age/NN wise/RB he's at the upper limit

- Collocations with *type* should also be done as separate words

a dinner/NN type/NN thing/NN
a polluting/NN type/NN deal/NN
a career/NN type/NN position/NN
P/NN C/NN type/NN things/NNS
sight/NN seeing/NN type/NN stuff/NN
an Austin/NNP Heely/NNP type/NN engine/NN

- Do not use GW to join parts of numbers either to other numbers of to anything else. Numbers are always labelled CD.

he could build a Seven/CD Forty-Seven/CD out of it
a/DT new/JJ four/CD eighty/CD six/CD chip/NN
big/JJ old/JJ four-/CD foot/NN center/NN board/NN
four/CD wheel/NN drive/NN
a/DT catch/NN twenty-two/CD
the/DT ninety/CD eighth/JJ person/NN
Route/NNP Seventy/CD

- Collocations with *free* should be done as separate words

  lint/NN free/JJ
  violence/NN free/JJ

- Do not use GW to join noun-noun collocations together.

  rain/NN storms/NNS
  bus/NN driver/NN
  a/DT computer/NN science/NN degree/NN
  head/NN phones/NNS
  back/NN muscles/NNS
  water/NN aerobics/NNS
  air/NN conditioning/NN
  living/NN room/NN
  dining/NN room/NN
  a tape/NN recording/NN
  waiting/NN list/NN
  bunk/NN bed/NN
  sweat/NN shirt/NN (BUT T/GW shirt/NN)

- Label each part of proper names.

  a Mud/NNP Hens/NNPS double/JJ header/NN
  a Sky/NNP Hawk/NNP Buick/NNP
  a leading/NNP edge/NNP
  (this is the name of a computer, which the transcriber clearly didn't know)

- Don't use GW on adjective-noun collocations, even if they seem closely tied.

  hot/JJ tub/NN
  cool/JJ whip/NN

- Do not use GW to join together foreign phrases. Just label both parts FW.

  per/FW capita/FW
  per/FW se/FW
  et/FW cetera/FW

- Nouns derived from verb-particle collocations (*dropout, breakdown*, etc.) are considered single words (see above). But when used adjectivally or in the passive, these are tagged separately.

  the cut/VBN up/RP tomatos/NNS
  a built/VBN in/RP window/NN box/NN
  this beaten/VBN down/RP path/NN
  they even have dress/VB up/RP days/NNS
  the teachers I know wear dress/VB up/RP jeans/NNS
  a call/VB in/RP survey/NN

a fix/VB up/RP special/NN
pull/VB out/RP couches/NNS
it had some well/RB thought/VBN out/RP parts/NN
watered/VBN down/RP wine/NN

he got himself really messed/JJ up/RP
he just got fed/JJ up/RP
I'm just kind of spaced/JJ out/RP
we're so wound/JJ up/RP in the boy scouts
I just feel hemmed/JJ in/RP by that
everyone is spread/JJ out/RP all over TImbuktu
I don't typically feel intruded/JJ on/RP
their home is paid/JJ for/RP
The pages were torn/VBN out/RP by my little sister

- Multi-word modifiers of any type should be tagged word by word (unless Noah gives them as hyphenated, in which case you MAY, but don't have to use GW)

  well/RB dressed/VBN politicians

  defense/NN oriented/VBN military/NN
  agression/NN oriented/VBN militaryNN
  a state/NN sponsored/VBN school/NN

  fast/RB breaking/VBG events/NNS
  the eye/NN stinging/VBG variety/NN
  time/NN consuming/VBG projects/NNS
  problem/NN solving/VBG skills/NNS

  the wood is very slow/RB burning/JJ

  sloppy/JJ looking/JJ jeans/NNS
  funny/JJ looking/JJ coke/NN
  strange/JJ looking/JJ can/NN
  country/NN looking/JJ watermelon/NN

  blond/JJ headed/JJ girl/NN
  open/JJ minded/JJ person/NN
  liberal/JJ minded/JJ politician/NN

  the/DT lower/JJR end/NN of/IN the/DT top/NN of/IN the/DT line/NN hotels/NNS
  break/NN for/IN out/IN of/IN state/NN students/NNS
  an/DT up/RB and/CC coming/VBG team/NN
  out/IN of/IN body/NN experiences/NNS
  he was on a/DT year/NN and/CC a/DT half/JJ training/NN plan/NN

- Do *a while* as separate *a/DT while/NN*, although Noah allows for both *a while* and *awhile*.

3 Typo Policy

Typos are indicated by a caret (ˆ) preceding the tag. The tag given is the tag for the hypothesized correct word, not the actual word.

Words which are wrongly capitalized (or wrongly not capitalized) should not on that account alone count as typos. They should be tagged as nouns or proper nouns according to sense not capitalization. Proper names which are wrongly spelled may be labelled typos but it is not necessary.

The typos exemplified here are common and fairly certain. There are many other cases which are more shaky. As a general rule, something should only count as a typo if it is a homophone or a near homophone (defined as not more than one sound (vowel or consonant) different), but you'll have to use your judgement about individual cases. There are some examples of things that are definitely not typos at the end of this message.

```
A.1 Homophones

When the typo is a homophone of the correct word it is tagged with the typo
sign (^) and the tag for the correct word.

put/VB there/^PRP$ money/NN in/IN places/NNS (=their)

it/PRP 's/BES really/RB not/RB to/^RB bad/JJ (=too)

their/PRP$ reporters/NNS needed/VBN to/^CD shots/NNS (=two)

know/^DT  matter where you build it, (=no)

right/^VB a book about it (=write)

I/PRP get/VBP to/TO here/^VB about/IN Texas/NNP (=hear)

fate through/^VBD them together /  (=threw)

you/PRP could/MD say/VB high/^UH to/TO your/PRP$ teachers/NNS (=hi)

He one/^VBD the race (=won)

I/PRP might/MD of/^VB ./. (=have)


A.2 Semi-homophones (or homophones in some dialects)

Some words that are not homophones in standard English are in speech or at
least in some dialects. Some of these are pretty standard (like final t/d
deletion) while others (higher=here) are heavy dialect. Some common examples
are listed below.

(a) final t/d deletion

everybody/NN is/NNS suppose/^VBN to/TO bring/VB something/NN
```

I/PRP use/^VBD to/TO play/VB racquetball/NN --/:

I/PRP would/MD have/VB like/^VBN to/TO have/VB

ham and bake/^VBN potatoes,

(b) an/and confusion

An/^CC they/PRP have/VBP to/TO be/VB in/IN ideal/JJ physical/JJ shape/NN ,/,
basically/RB ./. (an=and)

there/EX 's/BES and/^DT old/JJ joke/NN about/IN (and=an)

(c) than/then confusion

other/JJ then/^IN that/IN (then=than)

more reluctant of letting my older children go baby-sit for her because I
didn't know her -- --/: then/^IN she/PRP was/VBD reluctant/JJ of/IN
letting/VBG strangers/NNS into/IN her/PRP$ house/NN (then=than)

(d) where/were/we're

the/DT fact/NN that/IN people/NNS where/^VBD having/VBG this/DT problem/NN

when/WRB you/PRP where/^VBD in/IN school/NN

there/EX where/^VBD large/JJ numbers/NNS

not were/^WRB we are.

and/CC were/^PRP^VBP officially/RB ,/, yeah/UH ,/, and/CC we/PRP 're/VBP
officially/RB in/IN a/DT state/NN of/IN emergency/NN ./.

And/CC were/^PRP^VBP here/RB

(e) I/a/uh

as/IN a/^PRP say/NN (a=I)

where/WRB a/^PRP do/NN n't/RB (a=I)

it/PRP takes/VBZ almost/RB six/CD months/NNS to/TO get/VB ,/, uh/^DT ,/,
handgun/NN permit/NN (uh=a)

0926
I/PRP think/VBP that/DT 's/BES I/^DT pretty/RB good/JJ idea/NN ./. (I=a)

(f) accept/except, access/excess

Accept/^IN for what we give to my daughter

I think other then/^IN accept/^IN on a commercial or on news coverage
(note also 'then' for 'than')

in/IN access/^NN of/IN that/DT now/RB so/IN it/PRP

(g) are/or

Are/^CC they put one on the parents

the/DT first/JJ four/CD ,/, five/CD years/NNS or/^VBP so/RB important/JJ

just because people or/^VBP so, I don't know, just today people are just so money hungry,

(h) i/e confusion

they/PRP did/VBD n't/RB have/VB a/DT since/^NN of/IN risk/NN (since=sense)

I well/^MD never, ever go back. (well=will)

had/VBD some/DT bends/^NNS (from context clearly = bins)

(i) s/c confusion

investment advise/^NN

somebody/NN trying/VBG to/TO ,/, to/TO device/^VB a/DT scam/NN

(i) miscellaneous (some of these are probably keyboard mistakes)

I/PRP 'd/MD a/^VB never/RB put/VBN (a=have)

Here/^PRP$ name/NN is/VBZ Lori/NNP (here=her)

no/DT skin/NN of/^IN my/PRP$ back/RB (of=off)

hugh/^JJ companies like Three M, uh, I'm, uh, Honeywell (hugh=huge)

our/PRP$ cancer/NN society/NN sales/^VBZ daffodils/NNS ./. (sales=sells)

I think it's four cups of floor/^NN , (floor=flour)

the/DT okra/NN that/WDT is/VBZ growing/VBG around/IN higher/JJR (higher=here?)

A.3 Keyboard mistakes

If a word has one wrong letter or is missing a letter and it is obvious from context what the word should be, count it as a typo. The following sets at least are allowable as typos.

(a) than/that

he's a better man that/^IN I am (that=than)

(b) if/it/is/in

in/^PRP was/VBD in/IN high/JJ school/NN (in=it)

somebody does if/^PRP and breaks tradition (if=it)

especially it/^IN he's a repeater (it=if)

Is/VBZ is/^PRP just/RB aerobics/NN or/CC ,/, (is=it (probably))

(c) on/of/or

take it out or your paycheck (or=of)

(d) to/do/so

So/RB ,/, what/WP else/RB to/^VBP you/PRP tape/VBP (to=do)

What/WDT sort/NN of/IN requirements/NNS to/^VBP you/PRP have/VBP (to=do)

just/RB so/^TO see/VB what/WP ,/, uh/UH ,/, they/PRP might/MD have/VB to/TO offer/VB (so=to)

push/VB it/PRP so/^IN the/DT end/NN --/: (so=to)

(e) out/our

great/JJ out/IN country/NN really/RB is/VBZ ./. (out=our)

(f) the/they

that/DT 's/BES what/WP the/^PRP prefer/VBP to/TO do/VB (the=they)

they/^DT one we was all worried about -- (they=the)

(g) miscellaneous (some might be dialect rather than keyboard mistakes)

I/PRP do/VBP n't/RB thing/^VB they/PRP (thing=think)

just/RB short/^RB of/IN have/VBP it/PRP on/IN (short=sort)

does/VBZ [you/^PRP$ husband/NN] (you=your)

I/PRP was/VBD just/RB wandering/VBG if/IN (wandering=wondering)

he's got now/^DT doubt that (now=no)

one/CD or/CC two/CD timer/^NNS a/DT year/NN (timer=times)

whether/IN I/PRP night/^MD ,/, uh/UH ,/, uh/UH ,/, (night=might)

we/PRP 're/VBP taking/^VBG about/IN monies/NNS way/RB (taking=talking)

B. Words that should be split but aren't

A word that should be split is given two tags, one for each of the parts. Each of the tags is preceded by the typo sign (^)

(a) cliticized verbs not separated

If/IN your/^PRP^VBP happy/JJ with/IN it/PRP (=you 're)

their/^PRP^VBP offering/VBG a/DT service/NN (=they 're)

whose/^WP^BES going/VBG to/TO really/RB make/VB them/PRP ./. (=who 's (who is))

Someone whose/^WP^HVS got accounts (=who 's (who has))

Okay/UH ,/, I/PRP guess/VBP its/^PRP^BES recording/NN ./. (=it 's)

my/PRP$ husbands/^NN^BES retired/VBN ./. (=husband 's)

[The, + the] deposits/^NN^BES only on like drink stuff. (=deposit 's)

who/WP do/VBP you/PRP thinks/^VB^BES going/VBG to/TO win/VB (=think 's)

(b) cliticized pronoun not separated

lets/^VB^PRP turn/NN the/DT war/NN off/RP (=let 's)

(c) singular possessive /'s/ not separated

what/WP the/DT guys/^NN^POS name/VBP is/VBZ (=guy 's)

my/PRP$ in-laws/^NN^POS place/NN (=in-law 's)

my/PRP$ neighbors/^NN^POS yard/NN ./. (=neighbor 's)

(d) plural possessive apostraphe missing (if it were present it would be tagged POS)

your cats/^NNS^POS names (=cats ')

the students/^NNS^POS , uh, parents (=students ')

(e) verb-particle combinations (but NOT noun-particle combinations, which should be joined, and must be joined up using GW if they are separated)

people/NNS giveaway/^VBP^RB personal/NN information/NN (=give away)

It/PRP was/VBD already/RB setup/^VBN^PRT (=set up)

If he doesn't back-up/^VB^PRT (=back up)

(f) other miscellaneous cases

for/IN along/^DT^JJ time/NN (=a long)

for/IN awhile/^DT^NN ./. (=a while)

want/VB to/TO have/VB anymore/^DT^JJ children/NNS (=any more)

that/DT maybe/^MD^VB true/JJ (=may be)

we went fishing everyday/^DT^NN (=every day)

C. Words that are separate but shouldn't be

Parts of words that are separated are joined with the GW tag. The final part carries the label for all the joined parts (not usually more than two) and it has the typo sign preceding its label. Note that the typo sign is used on the final part even if it would have the right tag anyway (see the 'with out' example below).

(a) plural /s/ written as possessive

other/JJ than/IN just/RB it/GW 's/^PRP$ labor/NN (=its)

the/DT Honda/GW 's/^NNPS have/VBP been/VBN very/RB safe/JJ ./. (=Hondas)

I wouldn't be surprised if thing/GW 's/^NNS like that didn't happen (=things)

our/PRP$ causes/NNS do/VBP not/RB seem/VB as/RB important/JJ as/IN you/PRP
all/GW 's/^PRP were/VBD (=you alls) ??

(b) third singular /s/ written as possessive

Well/UH ,/, what/WP get/GW 's/^VBZ me/PRP is/VBZ (=gets)

(c) they're = their, you're = your

they/GW 're/^PRP$ Mom would never know it. (=their)

I gave him you/GW 're/^PRP$ book (=your)

(b) 'a-' words

well/UH ,/, let/VB me/PRP go/VB a/GW head/^RB (=ahead)

they don't look a thing a/GW like/^JJ . (=alike)

There's a certain a/GW mount/^NN of dribble (=amount)

(c) 'all-' words

all/GW though/^IN I have a feeling that people look (=although)

It/PRP 's/BES all/DT metric/JJ all/GW ready/^RB ./. (=already)

(d) miscellaneous examples

a/DT meal/NN with/GW out/^IN any/DT vegetables/NNS (=without)

I don't like him any/GW more/RB (=anymore)

D. Definitely not typos

(a) numbers in words indicate accents, just ignore them

,/, and/CC my/PRP$ husband/NN ,/, then/RB fianc3e/NN

(b) don't correct people's grammar, if they use the wrong tense of a verb or

whatever, just leave it

latest/JJS one/NN I/PRP 've/VBP saw/VBD

(c) 'zero' spelled 'oh' is a CD and is not a typo

the/DT first/JJ appearance/NN of/IN Roger/NNP Moore/NNP as/IN double/JJ oh/CD seven/CD

we/PRP actually/RB do/VBP have/VB some/DT money/NN in/IN a/DT Four/CD Oh/CD
One/CD K/SYM

BEVERLY/NNP HILLS/NNP NINE/CD OH/CD TWO/CD ONE/CD OH/CD ,/,