# Annotation Guidelines for Twitter Part-of-Speech Tagging
# Version 0.3 (March 2013)

**Kevin Gimpel  Nathan Schneider  Brendan O'Connor**

Data and references are available at: `http://www.ark.cs.cmu.edu/TweetNLP/`

## 1 Introduction

Online conversational text differs markedly from traditional written genres like newswire, including in ways that bear on linguistic analysis schemes. Here we describe part-of-speech (POS) annotation guidelines for online conversational text, using the Twitter POS tagset from Gimpel et al. (2011). That paper introduced a tagset and presented experimental results for a supervised tagger trained on manually-annotated tweets, but no explicit tagging guidelines were presented.

In this document we provide the set of annotation guidelines that were used for manually tagging the new DAILY547 Twitter data set (as discussed in section 5 of Owoputi et al. (2013)) and for fixing inconsistencies Gimpel et al.'s original data set. The new dataset has been released as version 0.3 at:
`http://www.ark.cs.cmu.edu/TweetNLP/`

The 25 tags are:

- Nominal
  - **N** – common noun
  - **O** – pronoun (personal/WH; not possessive)
  - **^** – proper noun
  - **S** – nominal + possessive
  - **Z** – proper noun + possessive
- Other open-class words
  - **V** – verb incl. copula, auxiliaries
  - **A** – adjective
  - **R** – adverb
  - **!** – interjection
- Other closed-class words
  - **D** – determiner
  - **P** – pre- or postposition, or subordinating conjunction
  - **&** – coordinating conjunction
  - **T** – verb particle
  - **X** – existential *there*, predeterminers
- Twitter/online-specific
  - **#** – hashtag (indicates topic/category for tweet)
  - **@** – at-mention (indicates another user as a recipient of a tweet)
  - **~** – discourse marker, indications of continuation of a message across multiple tweets
  - **U** – URL or email address
  - **E** – emoticon
- Miscellaneous
  - **$** – numeral
  - **,** – punctuation
  - **G** – other abbreviations, foreign words, possessive endings, symbols, garbage
- Other compounds
  - **L** – nominal + verbal (e.g. *i'm*), verbal + nominal (*let's*, *lemme*)
  - **M** – proper noun + verbal
  - **Y** – X + verbal

Compound tags are used because tokenization is difficult and sometimes counterproductive for the nonstandard orthographic forms that are commonplace in online conversational text; see section 5.1 of Owoputi et al. (2013). We discuss challenges of dealing with tokenization in the next section and then continue our discussion of POS annotation in the remaining sections.

## 2 Tokenization

When multiple words are written together without spaces, or when there are spaces between all characters of a word, or when the tokenizer fails to split words, the resulting tokens are tagged as 'G'. This is illustrated by the bold tokens in the following examples:

> It's Been Coldd :/ **iGuess** It's Better Than Beingg Hot Tho . Where Do Yuhh Live At ? @anonuser
>
> This yearâs almond crop is a great one . And the crop is being shipped fresh to **youâŠNow** !
>
> RT @anonuser Um ..... #TeamHeat ..... wut Happend ?? Haha <== #TeamCeltics showin that ass what's **good...That's** wat happened !!! LMAO
>
> #uWasCoolUntil you unfollowed me ! **R E T W E E T** if you Hate when people do that for no reason .

## 3 Penn Treebank Conventions

Generally, we followed the Penn Treebank (PTB; Marcus et al., 1993) Wall Street Journal conventions in determining parts of speech. However, there are many inconsistencies in the PTB annotations. We attempted to follow the majority convention for a particular use of a word, but in some cases we did not. Specific cases which caused difficulty for annotators or necessitated a departure from the PTB approach are discussed below.

### 3.1 Gerunds and Participles

Verb forms are nearly always given a verbal tag in the PTB (VBG for *-ing* forms, VBD for *-ed* forms), so we generally tag them as 'V' in our dataset. However, PTB sometimes tags as nouns or adjectives words such as *upcoming*, *annoying*, *amazing*, *scared*, *related* (as in *the related article*), and *unexpected*. We follow the PTB in tagging these as adjectives or nouns, appropriately.

### 3.2 Numbers and Values

- **Cardinal numbers** are tagged as '$'.
- **Ordinal numbers** are typically tagged as adjectives, following the PTB, except for cases like *28th October*, in which *28th* is tagged as '$'.
- **Times**: Following the Treebank, *A.M.* and *P.M.* are common nouns, while time zones (*EST*, etc.) are proper nouns.
- **Days, months, and seasons**: Like the PTB, days of the week and months of the year are always tagged as proper nouns, while seasons are common nouns.
- **Street addresses**: We follow the PTB convention of tagging numbers (house numbers, street numbers, and zip codes) as '$' and all other words in the address as proper nouns, like in the following PTB example: *153/CD East/NNP 53rd/CD St./NNP*.

  However, this convention is not entirely consistent in the PTB. Certain street numbers in the PTB are tagged as proper nouns: *Fifth/NNP Ave/NNP*

  Annotators are to use their best judgment in tagging street numbers.

- **Cardinal directions**: tokens such as *east* and *NNW* when referred to in isolation (not as a modifier or part of a name) are tagged as common nouns.
- **Units of measurement** are common nouns, even if they come from a person's name (like *Celsius*).

### 3.3 Time and Location Nouns Modifying Verbs

In the PTB, time and location nouns (words like *yesterday*/*today*/*tomorrow*, *home*/*outside*, etc.) that modify verbs are inconsistently labeled. The words *yesterday*/*today*/*tomorrow* are nearly always tagged as nouns, even when modifying verbs. For example, in the PTB *today* is tagged as NN 336 times and RB once. We note, however, that sometimes the parse structure can be used to disambiguate the NN tags. When used as an adverb, *today* is often the sole child of an NP-TMP, e.g.,

```
(NP-SBJ (DT These) (JJ high) (NNS rollers) )
(VP (VBD took)
    (NP (DT a) (JJ big) (NN bath) )
    (NP-TMP (NN today) ))
```

When used as a noun, it is often the sole child of an NP, e.g.,

```
(PP-TMP (IN until) (NP (NN today) ))
```

Since we are not annotating parse structure, it is less clear what to do with our data. In attempting to be consistent with the PTB, we typically tagged *today* as a noun.

The PTB annotations are less clear for words like *home*. Of the 24 times that *home* appears as the sole child under a *DIR* (direction) nonterminal, it is annotated as:

- `(ADVP-DIR (NN home) )` 14 times
- `(ADVP-DIR (RB home) )` 6 times
- `(NP-DIR (NN home) )` 3 times
- `(NP-DIR (RB home) )` 1 time

Manual inspection of the 24 occurrences revealed no discernible difference in usage that would warrant these differences in annotation. As a result of these inconsistencies, we decided to let annotators use their best judgment when annotating these types of words in tweets, asking them to refer to the PTB and to previously-annotated data to improve consistency.

### 3.4 Names

In general, every noun within a proper name should be tagged as a proper noun ('^'):

- Jesse/^ and/& the/D Rippers/^
- the/D California/^ Chamber/^ of/P Commerce/^

Company and web site names (*Twitter*, *Yahoo ! News*) are tagged as proper nouns. Function words are only ever tagged as proper nouns if they are not behaving in a normal syntactic fashion, e.g. *Ace/^of/^Base/^*.

**Personal names:** Titles/forms of address with personal names should be tagged as proper nouns: *Mr.*, *Mrs.*, *Sir*, *Aunt*, *President*, *Captain*. On their own—not preceding someone's given name or surname—they are common nouns, even in the official name of an office: *President/^Obama/^said/V*, *the/D president/N said/V*, *he/O is/V president/N of/P the/D U.S./^*

**Titles of works:** In the PTB, simple titles like *Star Wars* and *Cheers* are tagged as proper nouns, but titles with more extensive phrasal structure are tagged as ordinary phrases, e.g., *A/DT Fish/NN Called/VBN Wanda/NNP*. Note that *Fish* is tagged as NN (common noun). Therefore, we adopt the following rule: titles containing only nouns should be tagged as proper nouns, and other titles as ordinary phrases.

### 3.5 Prepositions and Particles

To differentiate between prepositions and verb particles (e.g., *out* in *take out*), we asked annotators to use the following test:

If you can insert an adverb within the phrasal verb, it's probably a preposition rather than a particle:

- turn slowly into/P a monster
- *take slowly out/T the trash

Below are other examples of verb particles:

- what's going on/T
- check it out/T
- shout out/T

(Relatedly, abbreviations like *s/o* and *SO* are tagged as 'V'.)

### 3.6 *this* and *that*: Demonstratives and Relativizers

The PTB almost always tags demonstrative *this*/*that* as a determiner, but in cases where it is used pronominally, it is immediately dominated by a singleton NP, e.g.

```
(NP (DT This)) is Japan
```

For our purposes, since we do not have parse trees and want to straightforwardly use the tags in POS patterns, we tag such cases as pronouns: *i just orgasmed over* ***this/O*** as opposed to ***this/D*** *wind is serious*. Words where we were careful about the 'D'/'O' distinction include, but are not limited to: *that, this, these, those, dat, daht, dis, tht*.

When *this* or *that* is used as a relativizer, we tag it as 'P' (never 'O'):

- You should know , **that/P** if you come any closer ...
- Never cheat on a woman **that/P** holds you down

The original Gimpel et al. (2011) data often used *this/D* for nominal usage, but was somewhat inconsistent. We changed the tags in their data to conform to the new style of these guidelines; all DAILY547 tags conform as well.

WH-word relativizers are treated differently than the above: they are sometimes tagged as 'O', sometimes as 'D', but never as 'P'.

### 3.7 Quantifiers and Referentiality

- A few non-prenominal cases of *some* are tagged as pronouns (*get some/O*). However, we use *some/D of*, *any/D of*, and *all/D of*.
- *someone*, *everyone*, *anyone*, *somebody*, *everybody*, *anybody*, *nobody*, *something*, *everything*, *anything*, and *nothing* are almost always tagged as nouns.

- *one* is usually tagged as a number, but occasionally as a noun or pronoun when it is referential, although this is inconsistent in the PTB and also in the annotated Twitter data.
- *none* is tagged as a noun.
- *all* and *any* are almost always tagged as a (pre)determiner or adverb.
- *few* and *several* are tagged as an adjective when used as a modifier, and as a noun elsewhere.
- *many* is tagged as an adjective.
- *lot* and *lots* (meaning a large amount/degree of something) are tagged as nouns.

### 3.8 Metalinguistic Mentions

Mentions of a word (typically in quotes) are tagged as if the word had been used normally:

- RT @anonuser Every girl lives for the " unexpected hugs from behind " moments < I wouldn't say " **live** "... but they r nice

Here *live* is tagged as a verb.

## 4 Phenomena in Twitter

There are many categories of phenomena that are frequent in Twitter and online conversational text that are not as frequent in the PTB. Below we describe how we treated them.

### 4.1 Hashtags and At-mentions

As discussed by Gimpel et al. (2011), **hashtags** used within a phrase or sentence are not distinguished from ordinary words. However, when the hashtag is external to the syntax and merely serves to categorize the tweet, it receives the '#' tag. **At-mentions** *always* receive the '@' tag, even though they occasionally double as words within a sentence.

### 4.2 Multiword Abbreviations

Some multiword abbreviations have natural tag correspondences: *lol* (laughing out loud) is typically an exclamation, tagged as '!'; *idk* or *iono* (I don't know) can be tagged as 'L' (nominal + verbal).

Miscellaneous types of abbreviations are tagged with 'G':

- *ily* (I love you)
- *wby* (what about you)
- *mfw* (my face when)

### 4.3 Clipping

Due to space constraints, words at the ends of tweets are sometimes cut off. We attempt to tag the partial word as if it had not been clipped. If the tag is unclear, we fall back to 'G':

- RT @anonuser : Tonight's memorial for Lucas Ransom starts at 8:00 p.m. and is being held at the open space at the corner of Del **Pla** ...

We infer that *Pla* is a clipped proper name, and accordingly tag it as '^'.

- RT @anonuser : Usually the people that need our help the most are the ones that are hardest 2 get through 2 . Be patient , love on **t** ...

The continuation is unclear, so we fall back to *t/G*.

### 4.4 Symbols, Arrows, etc.

- RT @anonuser : Helppp meeeee . I'mmm meltiiinngggg –> http://twitpic.com/316cjg
- http://bit.ly/cLRm23 <– #ICONLOUNGE 257 Trinity Ave , Downtown Atlanta ... Party This Wednesday ! RT

These arrows (–> and <–) are tagged as 'G'. But see the next section for Twitter discourse uses of arrows, which receive the '~' tag.

### 4.5 Twitter Discourse Tokens: Retweets, Continuation Markers, and Arrow Deixis

A common phenomenon in Twitter is the **retweet construction**, shown in the following example:

- RT @anonuser : Miami put a fork in it ...

The *RT* indicates that what follows is a "retweet" of another tweet. Typically it is followed by a Twitter username in the form of an at-mention followed by a colon (:). In this construction, we tag both the *RT* and *:* as '~'.

It is often the case that, due to the presence of the retweet header information, there is not enough space for the entirety of the original tweet:

- RT @anonuser : Because of the crisis in Haiti , I must now turn down the volume . We are one people . Let us find a way to show our huma ...

Here, the final ... is also tagged as '^' because it is not intentional punctuation but rather indicates that the tweet has been cut short due to space limitations (cf. "Clipping" above).

Aside from retweets, a common phenomenon in tweets is posting a link to a news story or other item of interest on the Internet. Typically the headline/title and beginning of the article begins the tweet, followed by ... and the URL:

- New NC Highway Signs Welcome Motorists To " Most Military Friendly State ": New signs going up on the major highways ... http://bit.ly/cPNH6e

Since the ellipsis indicates that the text in the tweet is continued elsewhere (namely at the subsequent URL), we tag it as '~'.

Sometimes instead of ..., the token *cont* (short for "continued") is used to indicate continuation:

- I predict I won't win a single game I bet on . Got Cliff Lee today , so if he loses its on me RT @anonuser : Texas ( cont ) http://tl.gd/6meogh

Here, *cont* is tagged as '~' and the surrounding parentheses are tagged as punctuation.

Another use of '~' is for tokens that indicate that one part of a tweet is a response to another part, particularly when used in an RT construction. Consider:

- RT @anonuser : First time seeing Ye's film on VH1 ≪ -What do you think about it ?

The ≪ indicates that the text after it is a response to the text before it (Collister, 2012), and is therefore tagged with '~'.

### 4.6 Nonstandard Spellings

We aim to choose tags that reflect what is *meant* by a token, even in the face of typographic errors, spelling errors, or intentional nonstandard spellings. For instance, in

- I'm here are the game! But not with the tickets from your brother. Lol

it is assumed that *at* was intended instead of *are*, so *are* is labeled as a preposition ('P'). Likewise, missing or extraneous apostrophes (e.g., *your* clearly intended as "you are") should not influence the choice of tag.

### 4.7 Direct Address

Words such as *girl*, *miss*, *man*, and *dude* are often used vocatively in tweets directed to another person. They are tagged as nouns in such cases:

- RT @anonuser : Shout-out to @anonuser definition of lil webbies i-n-d-e-p-e-n-d-e-n-t –> do you know what means **man** ??? << Ayyye !
- I need to go home **man** . Got some thangs I wanna try . Lol .

On the other hand, when such words do not refer to an individual but provide general emphasis, they are tagged as interjections ('!'):

- RT @anonuser : . #walesaid dt @anonuser should go DOWNTOWN ! Lol !!.. #jammy –>~ LMAO !! **Man** BIANCA !!!
- * Bbm yawn face * **Man** that #napflow felt so refreshing .
- **Man** da #Lakers have a fucking all-star squad fuck wit it !!

## References


Lauren B. Collister. 2012. The discourse deictics ^ and <– in a World of Warcraft community. *Discourse, Context & Media*, 1(1):9–19, March.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proc. ACL*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. NAACL*.