The Penn Treebank

M. Covington 2009 Feb 12

The Penn Treebank is licensed data. (UGA's copy cost about $2400 and you are not permitted to make copies for use elsewhere.) We store it in \\AIHV\NLP\PENN TREEBANK 3. The TXT directory contains it in the form of normal Windows text files.

The Treebank was created at the University of Pennsylvania. Human experts were hired to tag and parse several large corpora of English text: the Brown Corpus of samples of a wide variety of English-language writing, a large corpus from the Wall Street Journal, and a large corpus of telephone conversations ("the Switchboard Corpus"), and a small set of airline-related queries.

The symbols used for syntactic categories are different than in most other work. For example, a noun is NN, not N. See chart on next page.

Here is part of one of the tagged files.
Noun phrases have been marked with [ ] but no other syntactic structure is shown.

```
[ They/PRP ]
have/VB also/RB led/VBN
[ the/DT nation/NN ]
in/IN
[ the/DT direction/NN ]
of/IN
[ a/DT welfare/NN state/NN ]
./.
```

Here is the same material, parsed:

```
( (S
    (NP-SBJ (PRP They) )
    (VP (VB have)
      (ADVP (RB also) )
      (VP (VBN led)
        (NP (DT the) (NN nation) )
        (PP (IN in)
          (NP
            (NP (DT the) (NN direction) )
            (PP (IN of)
              (NP (DT a) (NN welfare) (NN state) ))))))
    (. .) ))
```

Table 1: Penn Treebank tag set (Santorini 1995).

Tag	**Category**	**Example**
CC	Coordinating conjunction	*and, but*
CD	Cardinal number	*three*
DT	Determiner	*the, a*
EX	Existential *there*	*there (is...)*
FW	Foreign word	*château*
IN	Preposition (≠ *to*) or subordinating conjunction	*with, after, if*
JJ	Adjective	*big*
JJR	Adjective, comparative	*bigger*
JJS	Adjective, superlative	*biggest*
LS	List item marker	*3.*
MD	Modal auxiliary verb	*shall*
NN	Noun (common)	*dog*
NNP	Noun (proper)	*America*
NNPS	Noun (proper), plural	*Americans*
NNS	Noun (common), plural	*dogs*
PDT	Predeterminer	*all (the dogs)*
POS	Possessive ending	*'s, '*
PRP	Personal pronoun	*he, she, they, I*
PRP$	Possessive pronoun	*his, her, their, my*
RB	Adverb or degree word	*quickly, very, not*
RBR	Adverb, comparative	*faster*
RBS	Adverb, superlative	*fastest*
RB	Particle	*(look it) up*
SYM	Symbol or formula in text	
TO	*to* whether prep. or verb marker	*to*
UH	Interjection	*wow!*
VB	Verb, plain form, not present tense	*(will) go*
VBD	Verb, past tense	*went, departed*
VBG	Verb, -*ing* form	*going*
VBN	Verb, past participle	*gone, departed*
VBP	Verb, plain form, present tense	*(we) go*
VBZ	Verb, -*s* form	*goes*
WDT	*Wh*-determiner	*which*
WP	*Wh*-pronoun	*what, who whom*
WP$	Possessive *wh*-pronoun	*whose*
WRB	*Wh*-adverb	*where, why, how*

Note that VB and VBP are always identical in form, and VBD and VBN are identical in form if the verb is regular.

Chart by M. Covington, 2008.