Penn Treebank II Tags

Note: This information comes from "Bracketing Guidelines for Treebank II Style Penn Treebank Project" - part of the documentation that comes with the Penn Treebank.

Contents:

Bracket Labels

Clause Level

S - simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a *wh*-word and that does not exhibit subject-verb inversion.
SBAR - Clause introduced by a (possibly empty) subordinating conjunction.
SBARQ - Direct question introduced by a *wh*-word or a *wh*-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.
SINV - Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.
SQ - Inverted yes/no question, or main clause of a *wh*-question, following the *wh*-phrase in SBARQ.

Phrase Level

ADJP - Adjective Phrase.
ADVP - Adverb Phrase.
CONJP - Conjunction Phrase.
FRAG - Fragment.
INTJ - Interjection. Corresponds approximately to the part-of-speech tag UH.
LST - List marker. Includes surrounding punctuation.
NAC - Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.
NP - Noun Phrase.
NX - Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.
PP - Prepositional Phrase.
PRN - Parenthetical.
PRT - Particle. Category for words that should be tagged RP.
QP - Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.
RRC - Reduced Relative Clause.
UCP - Unlike Coordinated Phrase.

VP - Vereb Phrase.
WHADJP - *Wh*-adjective Phrase. Adjectival phrase containing a *wh*-adverb, as in *how hot*.
WHAVP - *Wh*-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a *wh*-adverb such as *how* or *why*.
WHNP - *Wh*-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some *wh*-word, e.g. *who*, *which book*, *whose daughter*, *none of which*, or *how many leopards*.
WHPP - *Wh*-prepositional Phrase. Prepositional phrase containing a *wh*-noun phrase (such as *of which* or *by whose authority*) that either introduces a PP gap or is contained by a WHNP.
X - Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing *the...the*-constructions.

Word level

CC - Coordinating conjunction
CD - Cardinal number
DT - Determiner
EX - Existential there
FW - Foreign word
IN - Preposition or subordinating conjunction
JJ - Adjective
JJR - Adjective, comparative
JJS - Adjective, superlative
LS - List item marker
MD - Modal
NN - Noun, singular or mass
NNS - Noun, plural
NNP - Proper noun, singular
NNPS - Proper noun, plural
PDT - Predeterminer
POS - Possessive ending
PRP - Personal pronoun
PRP$ - Possessive pronoun (prolog version PRP-S)
RB - Adverb
RBR - Adverb, comparative
RBS - Adverb, superlative
RP - Particle
SYM - Symbol
TO - to
UH - Interjection
VB - Verb, base form
VBD - Verb, past tense
VBG - Verb, gerund or present participle
VBN - Verb, past participle
VBP - Verb, non-3rd person singular present
VBZ - Verb, 3rd person singular present
WDT - Wh-determiner
WP - Wh-pronoun
WP$ - Possessive wh-pronoun (prolog version WP-S)
WRB - Wh-adverb

Function tags

Form/function discrepancies

-ADV (adverbial) - marks a constituent other than ADVP or PP when it is used adverbially (e.g. NPs or free ("headless" relatives). However, constituents that themselves are modifying an ADVP generally do not get -ADV. If a more specific tag is available (for example, -TMP) then it is used alone and -ADV is implied. See the Adverbials section.
-NOM (nominal) - marks free ("headless") relatives and gerunds when they act nominally.

Grammatical role

-DTV (dative) - marks the dative object in the unshifted form of the double object construction. If the preposition introducing the "dative" object is *for*, it is considered benefactive (-BNF). -DTV (and -BNF) is only used after verbs that can undergo dative shift.
-LGS (logical subject) - is used to mark the logical subject in passives. It attaches to the NP object of *by* and not to the PP node itself.
-PRD (predicate) - marks any predicate that is not VP. In the *do so* construction, the *so* is annotated as a predicate.
-PUT - marks the locative complement of *put*.
-SBJ (surface subject) - marks the structural surface subject of both matrix and embedded clauses, including those with null subjects.
-TPC ("topicalized") - marks elements that appear before the subject in a declarative sentence, but in two cases only:

1. if the front element is associated with a *T* in the position of the gap.
2. if the fronted element is left-dislocated (i.e. it is associated with a resumptive pronoun in the position of the gap).

-VOC (vocative) - marks nouns of address, regardless of their position in the sentence. It is not coindexed to the subject and not get -TPC when it is sentence-initial.

Adverbials

Adverbials are generally VP adjuncts.

-BNF (benefactive) - marks the beneficiary of an action (attaches to NP or PP).
This tag is used *only* when (1) the verb can undergo dative shift and (2) the prepositional variant (with the same meaning) uses *for*. The prepositional objects of dative-shifting verbs with other prepositions than *for* (such as *to* or *of*) are annotated -DTV.
-DIR (direction) - marks adverbials that answer the questions "from where?" and "to where?" It implies motion, which can be metaphorical as in *"...rose 5 pts. to 57-1/2"* or *"increased 70% to 5.8 billion yen"* -DIR is most often used with verbs of motion/transit and financial verbs.
-EXT (extent) - marks adverbial phrases that describe the spatial extent of an activity. -EXT was incorporated primarily for cases of movement in financial space, but is also used in analogous situations elsewhere. Obligatory complements do not receive -EXT. Words such as *fully* and *completely* are absolutes and do **not** receive -EXT.
-LOC (locative) - marks adverbials that indicate place/setting of the event. -LOC may also indicate metaphorical location. There is likely to be some varation in the use of -LOC due to differing annotator interpretations. In cases where the annotator is faced with a choice between -LOC or -TMP, the default is

-LOC. In cases involving SBAR, SBAR should not receive -LOC. -LOC has some uses that are not adverbial, such as with place names that are adjoined to other NPs and NAC-LOC premodifiers of NPs. The special tag -PUT is used for the locative argument of *put*.
-MNR (manner) - marks adverbials that indicate manner, including instrument phrases.
-PRP (purpose or reason) - marks purpose or reason clauses and PPs.
-TMP (temporal) - marks temporal or aspectual adverbials that answer the questions *when, how often*, or *how long*. It has some uses that are not strictly adverbial, auch as with dates that modify other NPs at S- or VP-level. In cases of apposition involving SBAR, the SBAR should not be labeled -TMP. Only in "financialspeak," and only when the dominating PP is a PP-DIR, may temporal modifiers be put at PP object level. Note that -TMP is not used in possessive phrases.

Miscellaneous

-CLR (closely related) - marks constituents that occupy some middle ground between arguments and adjunct of the verb phrase. These roughly correspond to "predication adjuncts", prepositional ditransitives, and some "phrasel verbs". Although constituents marked with -CLR are not strictly speaking complements, they are treated as complements whenever it makes a bracketing difference. The precise meaning of -CLR depends somewhat on the category of the phrase.

- **on S or SBAR** - These categories are usually arguments, so the -CLR tag indicates that the clause is more adverbial than normal clausal arguments. The most common case is the infinitival semi-complement of *use*, but there are a variety of other cases.
- **on PP, ADVP, SBAR-PRP, etc** - On categories that are ordinarily interpreted as (adjunct) adverbials, -CLR indicates a somewhat closer relationship to the verb. For example:
    - Prepositional Ditransitives
      In order to ensure consistency, the Treebank recognizes only a limited class of verbs that take more than one complement (-DTV and -PUT and Small Clauses) Verbs that fall outside these classes (including most of the prepositional ditransitive verbs in class [D2]) are often associated with -CLR.
    - Phrasal verbs
      Phrasal verbs are also annotated with -CLR or a combination of -PRT and PP-CLR. Words that are considered borderline between particle and adverb are often bracketed with ADVP-CLR.
    - Predication Adjuncts
      Many of Quirk's predication adjuncts are annotated with -CLR.
- **on NP** - To the extent that -CLR is used on NPs, it indicates that the NP is part of some kind of "fixed phrase" or expression, such as *take care of*. Variation is more likely for NPs than for other uses of -CLR.

-CLF (cleft) - marks it-clefts ("true clefts") and may be added to the labels S, SINV, or SQ.
-HLN (headline) - marks headlines and datelines. Note that headlines and datelines always constitute a unit of text that is structurally independent from the following sentence.
-TTL (title) - is attached to the top node of a title when this title appears inside running text. -TTL implies -NOM. The internal structure of the title is bracketed as usual.

Index of All Tags

CC

CD

-CLF

-CLR

CONJP

-DIR

DT

-DTV

EX

-EXT

FRAG

FW

-HLN

IN

INTJ

JJ

JJR

JJS

-LGS

-LOC

LS

LST

MD

-MNR

NAC

NN

NNS

NNP

NNPS

-NOM

NP

NX

PDT

POS

PP

-PRD

PRN

PRP

-PRP

PRP$ or PRP-S

PRT

-PUT

QP

RB

RBR

RBS

RP

RRC

S

SBAR

SBARQ