

# English PropBank Annotation Guidelines

**Claire Bonial**

*bonial@colorado.edu*

**Julia Bonn**

*julia.bonn@colorado.edu*

**Kathryn Conger**

*kathryn.conger@colorado.edu*

**Jena Hwang**

*hwangd@colorado.edu*

**Martha Palmer**

*mpalmer@colorado.edu*

**Nicholas Reese**

*nicholas.reese@colorado.edu*

Center for Computational Language and Education Research  
Institute of Cognitive Science  
University of Colorado at Boulder

Based on original PropBank guidelines by Olga Babko-Malaya

July 14, 2015

# Contents

<b>1</b>	<b>Verb Annotation Instructions</b>	<b>3</b>
1.1	PropBank Annotation Goals . . . . .	3
1.2	Sense Annotation . . . . .	4
1.2.1	Frame Files . . . . .	4
1.2.2	ER: Error roleset . . . . .	6
1.2.3	IE: Idiomatic Expression roleset . . . . .	7
1.2.4	DP: Duplicate indicator roleset . . . . .	7
1.2.5	LV: Light Verb roleset . . . . .	8
1.2.6	What to do when there is no frame file . . . . .	8
1.3	Annotation of Numbered Arguments . . . . .	8
1.3.1	Choosing ARG0 versus ARG1 . . . . .	8
1.3.2	ARGA: Secondary Agent . . . . .	10
1.4	Annotation of Modifiers . . . . .	10
1.4.1	Comitatives (COM) . . . . .	11
1.4.2	Locatives (LOC) . . . . .	12
1.4.3	Directional (DIR) . . . . .	12
1.4.4	Goal (GOL) . . . . .	12
1.4.5	Manner (MNR) . . . . .	13
1.4.6	Temporal (TMP) . . . . .	14
1.4.7	Extent (EXT) . . . . .	14
1.4.8	Reciprocals (REC) . . . . .	15
1.4.9	Secondary Predication (PRD) . . . . .	16
1.4.10	Purpose Clauses (PRP) . . . . .	16
1.4.11	Cause Clauses (CAU) . . . . .	17
1.4.12	Discourse (DIS) . . . . .	17
1.4.13	Modals (MOD) . . . . .	19
1.4.14	Negation (NEG) . . . . .	19
1.4.15	Direct Speech (DSP) . . . . .	19
1.4.16	Adverbials (ADV) . . . . .	21
1.4.17	Adjectival (ADJ) . . . . .	21
1.4.18	Light Verb (LVB) . . . . .	21
1.4.19	Construction (CXN) . . . . .	22
1.5	Span of Annotation . . . . .	22
1.6	Where to Place Tags . . . . .	26
1.6.1	Exceptions to Normal Tag Placement . . . . .	26
1.7	Understanding and Annotating Null Elements in the Penn TreeBank . . . . .	27
1.7.1	Passive Sentences . . . . .	28
1.7.2	Fronted and Dislocated Arguments . . . . .	29
1.7.3	Questions and Wh-Phrases . . . . .	30
1.7.4	Interpret Constituent Here (ICH) Traces . . . . .	31

1.7.5	Right Node Raising (RNR) Traces . . . . .	32
1.7.6	It EXtraPosition (EXP) . . . . .	35
1.7.7	Other Traces . . . . .	37
1.8	Linking and Annotation of Null Elements . . . . .	38
1.8.1	Relative Clause Annotation . . . . .	38
1.8.2	Reduced Relative Annotation . . . . .	38
1.8.3	Concatenation of multiple nodes into one argument . . . . .	41
1.8.4	Special cases of topicalization . . . . .	42
1.9	Special Cases: small clauses and sentential complements . . . . .	43
1.10	Handling common features of spoken data . . . . .	45
1.10.1	Disfluencies and Edited Nodes . . . . .	45
1.10.2	Asides: PRN nodes . . . . .	46
<b>2</b>	<b>Light Verb Annotation</b>	<b>47</b>
2.1	Pass 1: Verb Pass . . . . .	47
2.2	Pass 2: Noun Pass . . . . .	50
2.3	Examples . . . . .	50
2.4	Tricky Cases . . . . .	52
<b>3</b>	<b>Noun Annotation Instructions</b>	<b>53</b>
3.1	Span of Annotation . . . . .	53
3.2	Annotation of Numbered Arguments . . . . .	55
3.3	Annotation of Modifiers . . . . .	55
3.3.1	Adjectival modifiers (ADJ) . . . . .	56
3.3.2	Secondary Predication modifiers (PRD) . . . . .	56
3.4	Exceptions to Annotation: Determiners and Other Noun Relations . . . . .	57
3.5	.YY Roleset . . . . .	57
<b>4</b>	<b>Adjectival Predicate Annotation Instructions</b>	<b>59</b>
4.1	Span of Annotation . . . . .	59
4.2	Annotation of Arguments . . . . .	59
4.3	Annotating Constructions . . . . .	59
4.3.1	Comparative Constructions . . . . .	60
4.3.2	Degree Construction . . . . .	61
4.3.3	The Xer the Yer Construction . . . . .	62
<b>A</b>	<b>Jubilee Hotkeys</b>	<b>63</b>
	<b>References</b>	<b>65</b>

# Chapter 1

## Verb Annotation Instructions

### 1.1 PropBank Annotation Goals

PropBank is a corpus in which the arguments of each predicate are annotated with their semantic roles in relation to the predicate (Palmer et al., 2005). Currently, all the PropBank annotations are done on top of the phrase structure annotation of the Penn TreeBank (Marcus et al., 1993). In addition to semantic role annotation, PropBank annotation requires the choice of a sense ID (also known as a *frameset* or *roleset* ID) for each predicate. Thus, for each verb in every tree (representing the phrase structure of the corresponding sentence), we create a PropBank instance that consists of the sense ID of the predicate (e.g. `run.02`) and its arguments labeled with semantic roles.

An important goal is to provide consistent argument labels across different syntactic realizations of the same verb, as in...

[*John*]<sub>ARG0</sub> *broke* [*the window*]<sub>ARG1</sub>  
[*The window*]<sub>ARG1</sub> *broke*

As this example shows, the arguments of the verbs are labeled as numbered arguments: ARG0, ARG1, ARG2, and so on. The argument structure of each predicate is outlined in the PropBank frame file for that predicate. The frame file gives both semantic and syntactic information about each sense of the predicate lemmas that have been encountered thus far in PropBank annotation. The frame file also denotes the correspondences between numbered arguments and semantic roles, as this is somewhat unique for each predicate. Numbered arguments reflect either the arguments that are required for the valency of a predicate (e.g., agent, patient, benefactive), or if not required, those that occur with high-frequency in actual usage. Although numbered arguments correspond to slightly different semantic roles given the usage of each predicate, in general numbered arguments correspond to the following semantic roles:

ARG0	agent	ARG3	starting point, benefactive, attribute
ARG1	patient	ARG4	ending point
ARG2	instrument, benefactive, attribute	ARGM	modifier

Table 1.1: List of arguments in PropBank

In addition to numbered arguments, another task of PropBank annotation involves assigning functional tags to all modifiers of the verb, such as manner (MNR), locative (LOC), temporal (TMP) and others:

*Mr. Bush met him privately, in the White House, on Thursday.*  
REL: met

ARG0: Mr. Bush  
ARG1: him  
ARGM-MNR: privately  
ARGM-LOC: in the White House  
ARGM-TMP: on Thursday

And, finally, PropBank annotation involves finding antecedents for empty arguments of the verbs, as illustrated below:

*I made a decision [\*PRO\*] to leave.*

The subject of the verb *leave* in this example is represented as an empty category [\*] in TreeBank. In PropBank, all empty categories which could be co-referred with a NP within the same sentence are linked in co-reference chains:

REL: leave  
ARG0: [\*PRO\*] \* [I]

Similarly, relativizers and their referents are linked in relative clause constructions, and traces are linked to their referents in reduced relative constructions. While these links were at one time created manually by the annotators, they are now added automatically in post-processing.

The annotation of this information creates a valuable corpus, which can be used as training data for a variety of natural language processing applications. Training data, essentially, is what computer scientists and computational linguists can use to ‘teach the computer’ about different aspects of human language. Once this information is processed, it can guide future decisions on how to categorize and/or label different features in novel utterances outside of the PropBank corpus. Parallel PropBank corpora currently exist or are underway for English, Chinese, Arabic and Hindi. As a whole, the PropBank corpus has the potential to assist in natural language processing applications such as machine translation, text editing, text summary and evaluation as well as question answering.

Thus, the main tasks of PropBank annotation are: argument labeling, annotation of modifiers, choosing a sense for the predicate, and creating links for empty categories, relative clauses, and reduced relatives. Each of these aspects of annotation are discussed in detail below. Although some detail is provided in each section on how to annotate appropriately using the annotation tool, Jubilee, complete guidelines on the use of this tool are provided in the Jubilee technical report: *Jubilee: Propbank Instance Editor Guideline (Version 2.1)* (Choi et al., 2009).

## 1.2 Sense Annotation

### 1.2.1 Frame Files

The argument labels for each predicate are specified in the frame files, which are available at <http://verbs.colorado.edu/propbank/framesets-english-aliases/> and are also displayed in the frameset view of the annotation tool, Jubilee (see Jubilee technical report (Choi et al., 2009) for further information). Frame files provide predicate-specific descriptions of semantic roles and illustrate these roles by providing examples.

Frame File for the verb *expect*:

Roles:

ARG0: expecter  
ARG1: thing expected

#### Example

*Portfolio managers expect further declines in interest rates.*

ARG0: Portfolio managers  
REL: expect  
ARG1: further declines in interest rates

For some predicates, it is impossible to provide one set of semantic roles for all senses of the word. For example, the two senses of the verb *leave* in the examples below take different arguments:

*Mary left the room*  
*Mary left her daughter-in-law her pearls in her will*

In such cases, frame files distinguish two or more verb senses, which are called framesets or rolesets (this term is interchangeable, depending on what language is being annotated), and define argument labels specific to each roleset:

Roleset **leave.01**, *the act of moving away from*:  
ARG0: entity leaving  
ARG1: place, person, or thing left  
ARG2: destination

Roleset **leave.02**, *give, bequeath*:  
ARG0: giver, doner  
ARG1: thing given  
ARG2: benefactive

As previously mentioned, frame files are also found in the frameset view (upper-right pane) of the annotation tool, Jubilee. Initially, the predicate lemma followed by **.XX** is displayed here to indicate that no sense has been chosen yet (e.g. **leave.XX**). Subsequent rolesets are numerically displayed, (e.g. **leave.01**, **leave.02**, **leave.03**, etc.). Listed above the argument structure for a given roleset are a list of predicates a roleset can be used for. Rolesets may only be used for predicates listed in this alias section with the corresponding part of speech. This was not always the case. In early stages of PropBank, annotation of separate parts of speech was done with separate frame files. However, in 2014-2015, the frame files across related parts of speech were “unified.” The result of this is one roleset can now correspond to different, albeit related, parts of speech. For example, the framefile for **depend.01** shows v-alias: *depend* and j-alias: *dependent*. This frame can be used for the *rely* sense of the verbal predicate *depend* and the adjectival predicate *dependent*. If the nominal *dependence* were used, annotators would need to fill out a problem report so the nominal senses could be added to the frame. See Section 1.2.6 for more information on what to do when there is no roleset.



Figure 1.1: Example of the Frameset View

When annotating, annotators first select the appropriate roleset (or sense), and then assign the argument labels as specified for this roleset. In Jubilee, annotators can peruse the available numbered senses of a predicate by clicking on the roleset combo-box, or they can move through

the available rolesets sequentially by using the shortcut `]` to move forward, or `[` to move back to lower-numbered senses. As a roleset is selected, the argument structure and a short definition of that sense, which are extracted from the corresponding frameset file (e.g. `leave.xml`), appear in the roleset information pane. To view annotation examples of the currently selected roleset, click **[Example]** button (`Ctrl+E`). Occasionally, viewing the frameset directly on the Propbank website is beneficial to an annotator, as it may have additional notes from adjudicators on distinguishing different senses of the verb. For example, the roleset for `leave.08` contains the note “Most of these probably appear as passives. It is unclear whether benefactives can occur with the active usage. All of the active usages I’ve seen that appear to include the benefactive should actually be tagged as `leave.01`, because the *over* is actually part of the location. ex: ‘*he had left the pistol over with the servants at the further side of the great tree.*’” This explanatory note is also available through the **[View Roleset Comments]** button or by pressing (`Ctrl+C`).

To accommodate verb particle constructions (e.g. *give up*), the frame file defines not only several senses of each verb, but also several predicates reflective of the verb’s associated verb particle constructions. If a verb has a particle (marked as `PRT` in TreeBank), then it is considered a different predicate, and may or may not have a different set of semantic roles. For example, the frame file for the verb *keep* defines three predicates: predicate *keep* (which has 6 rolesets), and predicates *keep up* and *keep on*, which each encompass 1 roleset respectively. The following example gives the definition of the predicate *keep up* and an example usage:

**Predicate: keep up:**

`keep.05`, *maintain one’s position*:

ARG0: maintainer of position

ARG1: relative to what

*John can’t keep up with Mary’s rapid mood swings.*

ARG0: John

ARGM-MOD: *ca*

ARGM-NEG: *n’t*

REL: *keep up*

ARG1: *with Mary’s rapid mood swings*

Note that the relation (`rel`) in PB annotation should include both the verb and the particle. Thus, the annotator must concatenate the particle to the original relation to form a single predicate lemma, annotated with the `rel` tag. To concatenate the particle, choose the particle node (as opposed to selecting just the particle itself) and type `Ctrl+Shift+.` The resulting `rel` annotation will reflect the locations of both the original predicate and the concatenated particle in the annotation view; for example, `9:0,8:0-rel`.

### 1.2.2 ER: Error roleset

All tokens marked as verbs in the TreeBank should be annotated in PropBank; however, rarely a token is marked as the `rel` that is not truly a verb and should not be annotated. Because the lines between parts of speech are often fuzzy, annotators should annotate all cases of gerunds and past participles, even if they seem adjectival or nominal in usage. However, if the token marked as a verb is not ever used grammatically as an active verb, then it should not be annotated and the `ER` roleset should be selected. For example, *collonaded* in ‘*The collonaded house...*’ has been marked as a verb in the past in the TreeBank, but a web search shows that there are no attested usages of *collonade* as an active verb; thus, this instance was treated as an error and marked as `ER`. This roleset should also be selected when a verb is being used prepositionally, and therefore heads a prepositional phrase in the TreeBank. For example, although *accord* can appear as an active verb, prepositional usages such as ‘*According to our sources...*’ should be marked as `ER`. Other examples of verbs that are often used prepositionally are *base* and *give*

in usages such as ‘*Based on current research...*’ and ‘*Given the situation...*’ In each of these cases, the annotator will notice that the verb syntactically heads a prepositional phrase. In general, **ER** should be selected for error cases where the **rel** is not a verb.

The **ER** roleset should also be selected in the case of a TreeBank error. There are many different kinds TreeBank errors, but the following is a common one. Occasionally, in noun annotation, there should be a **NML** node within an **NP** node, but the **NML** is missing. In some cases, the result is that constituents which should be grouped together within an **NML** would have to be tagged individually, requiring concatenation. This is not proper tagging procedure, and should not be done. In the following example, the **rel** should be contained in an **NML**, which it would be the head of. If this were done correctly, the instance would be taggable. Instead, it looks like the **rel** is not the head. This is a TreeBank error, and should be tagged as an **ER** roleset.

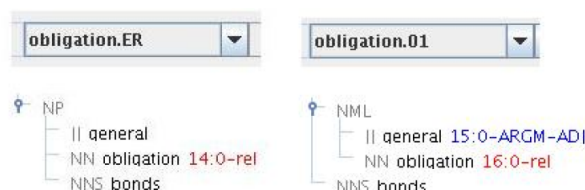


Figure 1.2: Example of a head which requires an **NML** node

### 1.2.3 IE: Idiomatic Expression roleset

If the **rel** is part of an idiomatic expression, wherein the meaning of the expression has no relationship to the meaning of its parts, then the **IE** roleset should be selected. Because PropBank has coarse-grained senses which treat metaphorical extensions of a sense in the same manner as the literal sense, annotators should be careful not to use **IE** where the idiomatic expression is metaphorically related to the words that comprise it. For example, *let* in ‘*let the cat out of the bag*’ should be annotated because ‘*the cat*’ is metaphorically related to a secret and ‘*the bag*’ is related to the secret’s hidden nature. However, *trip* in ‘*Trip the light fantastic toe*’ (meaning *to dance*), has no relationship to dancing, and similarly, *kick* in ‘*Kick the bucket*’ has no relationship to dying. In these cases, and only these cases, where the meaning of the expression is in no way related to the individual meanings of the words, the **IE** roleset should be selected.

### 1.2.4 DP: Duplicate indicator roleset

This roleset is not available to annotators, but is used in post-processing to indicate that a duplicate was necessary to handle a verb that has two separate argument structures. This occurs only in cases of verb ellipsis, for example, ‘*I ate a sandwich and Cindy a banana.*’ The TreeBank uses special ‘=’ notation and numbered indices to indicate that the second argument structure shares a verb given earlier. In these cases, annotators should select the appropriate numbered roleset, and should annotate only the first argument structure. Then, the annotator should fill out a problem report on the PropBank website (<http://verbs.colorado.edu/propbank/>), indicating that the instance needs a duplicate. During adjudication and post-processing, this duplicate is added and the second argument structure is annotated. The annotation of the second argument structure receives the **DP** roleset during post-processing so that it is clear that this instance is a duplicate.



### 1.2.5 LV: Light Verb roleset

This roleset is used to flag a verb’s usage as a light verb usage (e.g., *take a walk*, *have a drink*). See Chapter 2 for more details on light verb annotation.

### 1.2.6 What to do when there is no frame file

Occasionally, annotators will come across new verbs that do not have existing frame files. In these cases, firstly check to see if this is a mislemmatization of a verb that already has a frame file (e.g., a Britishism or misspelling). Use the frames listed on the website (<http://verbs.colorado.edu/propbank/framesets-english-aliases/>) to check this. If it is a mislemmatization, use the argument structure outlined in the existing frame file and take a note of the task and instance number of this problem, along with the correct roleset (e.g. `color.01`). Using this information, fill out a problem report on the PropBank website (<http://verbs.colorado.edu/propbank/>). If this is not the case, the annotator will have to do some research to determine what an appropriate argument structure would be. Brainstorm other verbs that have similar syntax and semantics, and see if any of those verbs already have frame files. Try to model your annotation after that frame file. Again, fill out a problem report for this instance, taking note of the task and instance number along with an outline of the numbered arguments used and their semantic role correspondences. Also, include the roleset that served as a basis of comparison if one was used.

If desired, the annotators can consult the Unified Verb Index: <https://verbs.colorado.edu/verb-index/>. This has links to existing VerbNet (VN) and FrameNet (FN) information on a particular predicate, which can be used to understand what are commonly thought of as a verb’s core arguments. Where possible, PropBank is mapped to VerbNet thematic roles. If a mapping is appropriate, a roleset’s VN class will be listed at the top of the roleset on the frames website. VerbNet uses more canonical thematic roles instead of numbered arguments that are unique to the predicate. Thus, the annotator can use this as a resource for brainstorming argument structures for new verbs as well as coming to a better understanding of existing frame files by examining another analysis of that verb’s thematic roles.

## 1.3 Annotation of Numbered Arguments

### 1.3.1 Choosing ARG0 versus ARG1

In most cases, choosing an argument label is straightforward, given the verb specific definition of this label in the frame files. However, in some cases, it may be somewhat ambiguous whether an argument should be annotated as ARG0 or ARG1; thus, the annotator must decide between these labels based on the following explanations of what generally characterizes ARG0 and ARG1.

The ARG0 label is assigned to arguments which are understood as agents, causers, or experiencers. The ARG1 label is usually assigned to the patient argument, i.e., the argument which undergoes the change of state or is being affected by the action.

ARG0 arguments (which correspond to external arguments in government and binding (GB) theory) are the subjects of transitive verbs and a class of intransitive verbs called unergatives. Semantically, external arguments have what Dowty (1991) called Proto-Agent properties, such as:

1. Volitional involvement in the event or state
2. Causing an event or change of state in another participant

3. Movement relative to the position of another participant (Dowty, 1991)

Internal arguments (labeled as ARG1) are the objects of transitive verbs and the subjects of intransitive verbs called unaccusatives. These arguments have Proto-Patient properties, which means that these arguments:

1. Undergo change of state
2. Are causally affected by another participant
3. Are stationary relative to movement of another participant (Dowty, 1991)

Whereas for many verbs, the choice between ARG0 or ARG1 does not present any difficulties, there is a class of intransitive verbs (known as verbs of variable behavior), where the argument can be tagged as either ARG0 or ARG1.

[*A bullet*]<sub>ARG1</sub> *landed at his feet*  
[*He*]<sub>ARG0</sub> *landed*

Arguments which are interpreted as agents should always be marked as ARG0, independent of whether they are also the ones which undergo the action. In general, if an argument satisfies two roles, the highest ranked argument label should be selected, where

$$ARG0 > ARG1 > ARG2 - 6 > ARG M$$

Given this rule, agents are ranked higher than patients. If an argument is both an agent and a patient, then ARG0 label should be selected. An example of this is the verb *meet*:

*John and Mary met at a cocktail party.*

ARG0: John and Mary

REL: met

ARGM-LOC: at a cocktail party

Another good example of this is the verb *move*, which can have both an ARG0 (i.e. the mover) and an ARG1 (i.e. the moved), though in some cases ARG0 and ARG1 are the same:

*Digital 's move into mainframes*

ARG0: Digital 's

REL: move

ARG2: into mainframes

Not all ARG0s are agentive, however. There are many inanimate as well as clausal arguments which are being marked as ARG0s. These arguments are usually the ones which cause an action or a change of state.

A notion which might be useful for selecting ARG0 arguments is the notion of internally caused as opposed to externally caused eventualities, as defined in Levin and Rapaport (1995). In internally-caused eventualities, some property inherent to the argument of the verb is responsible for bringing about the eventuality. For agentive verbs such as *play*, *speak*, or *work*, the inherent property responsible for the eventuality is the will or volition of the agent who performs the activity. However, an internally caused eventuality need not be agentive. For example, the verbs *blush* and *tremble* are not agentive, but they, nevertheless, can be considered to denote internally caused eventualities, because these eventualities arise from internal properties of the arguments, typically an emotional reaction. In contrast to internally caused verbs, verbs which are externally caused inherently imply the existence of an external cause with an immediate control over bringing about the eventuality denoted by the verb: an agent, and instrument, a natural force, or a circumstance. Thus something breaks because of the existence of some external cause; something does not break because of its own properties (Levin and Hovav,

1995). The difference between internal and external causation is important for distinguishing ARG0s and ARG1s: the arguments which are responsible for bringing out the eventuality are ARG0s, whereas those which undergo an externally caused event are ARG1s.

To sum up, ARG0 arguments are the arguments which cause the action denoted by the verb, either agentively or not, as well as those which are traditionally classified as experiencers, e.g., the arguments of stative verbs such as *love*, *hate*, *fear*. ARG1 arguments, on the other hand, are those that change due to external causation, as well as other types of patient-like arguments.

### 1.3.2 ARG0: Secondary Agent

In addition to argument numbers 1-6 and modifiers, the tag ‘ARG0’ is available. This should be used to annotate secondary agents. Predicates that often take secondary agents will have this specified in the roleset and a clarifying example will be provided (e.g. *trot.01*); however, it is possible albeit rare for other verbs to be characterized by a secondary agent. In general, the secondary agent tag will only be used when the argument structure outlined in the roleset indicates that a proto-agent role, such as ‘the trotter’<sub>ARG0</sub> for sense *trot.01*, is already fulfilled, yet there is another animate agent causing the event:

*John trotted his horse.*

ARGA: John

REL: trotted

ARG0: his horse

It is noted above that the use of ARG0 is rare in PropBank. A separate project, AMR (Barnarescu et al., 2013) which also uses PropBank frame files does not have an ARG0 tag. Because of the close affiliation between AMR and PropBank, and due to reframing already necessary to facilitate the combining, or unification, of etymologically related frame files across parts of speech, many frames previously containing ARG0 have been reworked to include an agentive ARG0 ‘causer.’ For example, the old *walk.01* frame was similar to that of *trot.01*, with an ARG0 ‘walker,’ an ARG1 ‘path walked,’ but no role for *John* in a sentence such as ‘*John walked his dog.*’ This frame has been changed and reannotated to include three arguments:

*walk.01*

ARG0-PAG: walker, thing in motion (vnrole: 51.3.2-theme)

ARG1-PPT: entity being walked, where there is an outside Agent (vnrole: 51.3.2-location)

ARG2-LOC: path or underspecified location walked

In the future, frames necessitating the use of ARG0 are likely to be reworked, making this tag unnecessary.

## 1.4 Annotation of Modifiers

The following types of modifiers are being used in PropBank:

COM: Comitative

LOC: Locative

DIR: Directional

GOL: Goal

MNR: Manner

TMP: Temporal

EXT: Extent

REC: Reciprocals

PRD: Secondary Predication

PRP: Purpose  
CAU: Cause  
DIS: Discourse  
ADV: Adverbials  
ADJ: Adjectival  
MOD: Modal  
NEG: Negation  
DSP: Direct Speech  
LVB: Light Verb  
CXN: Construction

Note: In the sections that follow describing each type of modifier, many real examples drawn from the corpus are used. As such, these examples contain null elements (\*, PRO) and traces (\*T\*). These null elements and traces often use indices, or numbers, to show the relationship between a null element and its referent. Thus, the null element or trace may have a number listed after it, and that number is listed again next to a word or phrase. This indicates that the two are coreferential. For more information on null elements, refer to Section 1.7 and Section 1.8.

#### 1.4.1 Comitatives (COM)

Comitative modifiers indicate who an action was done *with*. This can include people or organizations (entities that have characteristics of prototypical agents: animacy, volition) but excludes objects, which would be considered instrumental modifiers. Although the formal term for this modifier is ‘comitative,’ annotators can think of this argument as ‘companion:’ a companion to another in the action of the verb.

*I sang a song with my sister.*

ARG0: I  
REL: sang  
ARG1: a song  
ARGM-COM: with my sister

*The man joined the club with his friend.*

ARG0: The man  
REL: joined  
ARG1: the club  
ARGM-COM: with his friend

*I-1 got kicked [\*-1] out of the class with the bully.*

REL: kicked  
ARG1: [\*-1]  
ARGM-DIR: out of the class  
ARGM-COM: with the bully

*The next morning, with a police escort, busloads of executives and their wives raced to the Indianapolis Motor Speedway.*

ARGM-TMP: The next morning  
ARGM-COM: with a police escort  
ARG0: busloads of executives and their wives  
REL: raced  
ARG1: to the Indianapolis Motor Speedway

### 1.4.2 Locatives (LOC)

Locative modifiers indicate where some action takes place. The notion of a locative is not restricted to physical locations, but abstract locations are being marked as LOC as well; such as ‘[in his speech]-(LOC) he was talking about...’

*The percentage of lung cancer deaths among the workers at the West Groton , Mass. , paper factory appears [\*-1] to be the highest for [any asbestos workers]-1 studied [\*-1] in Western industrialized countries , he said [0] [\*T\*-2] .*

ARG1: [\*]

REL: studied

ARGM-LOC: in Western industrialized countries

*Areas of the factory [\*ICH\*-2] were particularly dusty where-1 [the crocidolite]-8 was used [\*-8] [\*T\*-1] .*

ARGM-LOC: [\*T\*-1]

ARG1: [\*-8]

REL: used

*In his ruling , Judge Curry added an additional \$ 55 million [\*U\*] to the commission ’s calculations.*

ARGM-LOC: In his ruling

ARG0: Judge Curry

REL: added

ARG1: an additional \$ 55 million [\*U\*]

ARG2: to the commission ’s calculations

### 1.4.3 Directional (DIR)

Directional modifiers show motion along some path. *Source* modifiers are also included in this category. However, if there is no clear path being followed, ARGM-LOC should be used instead. Thus, ‘walk along the road’ is a directional, but ‘walk around the countryside’ is a location. Directional modifiers are also used for some particles, as in ‘back up.’

*What sector is [\*T\*-46] stepping forward [\*-2] to pick up the slack ? ” he asked [\*T\*-1]*

ARG1: [\*T\*-46]

REL: stepping

ARGM-DIR: forward

ARGM-PRP: [\*-2] to pick up the slack

*That response annoyed Rep. Markey , House aides said [0] [\*T\*-1] , and the congressman snapped back that there had been enough studies of the issue and that it was time for action on the matter .*

ARG0: the congressman

REL: snapped

ARGM-DIR: back

ARG1: that there had been enough studies of the issue and that it was time for action on the matter

### 1.4.4 Goal (GOL)

This tag is for the goal of the action of the verb. This includes the final destination of motion verbs and benefactive arguments that receive something, or modifiers that indicate that the action of the verb was done for someone or something, or on their behalf:

*The child fed the cat for her mother.*

ARG0: the child

REL: fed

ARG1: the cat

ARGM-GOL: for her mother

*The couple translated for the Americans.*

ARG0: the couple

REL: translated

ARGM-GOL: for the Americans

ARGM-GOL should also be used for modifiers that indicate the final resting place or destination of motion or transfer verbs:

*Workers dumped large burlap sacks of the imported material into a huge bin , poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used [\*] [\*] to make filters.*

ARG0: Workers

REL: dumped

ARG1: large burlap sacks of the imported material

ARGM-GOL: into a huge bin

*We publicized to the masses our determination to fight against evil.*

ARG0: We

REL: publicized

ARGM-GOL: to the masses

ARG1: our determination to fight against evil

*The walls crumbled to the ground.*

ARG1: The walls

REL: crumbled

ARGM-GOL: to the ground

Be careful to distinguish instances like the above example (*The walls crumbled to the ground*) from secondary predication (e.g. *he bled to death*); the difference being that the above example involves motion and coming to rest.

Many motion verbs involving a change of state, such as *rise*, and *fall*, already have a numbered argument for this semantic role. Similarly, many transfer verbs, such as *give*, and *distribute*, already have a numbered argument for this role. In these cases, as in all situations where we have numbered arguments that are also encompassed by ARGMs, continue to prioritize the use of the numbered argument over that of the ARGM.

#### 1.4.5 Manner (MNR)

Manner modifiers specify how an action is performed. For example, *works well* is a manner. Manner tags should be used when an adverb could be an answer to a question starting with *How?*.

*Among 33 men who-4 [\*T\*-4] worked closely with the substance, 28 [\*ICH\*-1] have died – more than three times the expected number.*

ARG0: [\*T\*-4]

REL: worked

ARGM-MNR: closely

ARG1: with the substance

*Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used [\*] [\*] to make filters.*

ARG0: Workers

ARGM-MNR: mechanically

REL: mixed

ARG1: the dry fibers

ARGM-LOC: in a process used [\*] [\*] to make filters

In both of these sentences, the ARGM-MNR can answer the question ‘*How?*’:

Q: How did the 33 men work with the substance?

A: Closely.

Q: How did the workers mix the dry fibers?

A: Mechanically.

While many manner modifiers are adverbs, other phrases can also show manner. These answer the question of ‘*How?*’ and are often instruments.

*The plumber unclogged the sink with a drain snake.*

ARG0: The plumber

REL: unclogged

ARG1: the sink

ARGM-MNR: with a drain snake

#### 1.4.6 Temporal (TMP)

Temporal ARGMs show when an action took place, such as ‘*in 1987*,’ ‘*last Wednesday*,’ ‘*soon*,’ or ‘*immediately*.’ Also included in this category are adverbs of frequency (*often*, *always*, *sometimes*, with the exception of *never*, see Section 1.4.14 below), adverbs of duration (*for a year*, *in an year*), order (*first*), and repetition (*again*)

*[A form of asbestos]-2 once used [\*-2] [\*] to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed [\*] to it more than 30 years ago , researchers reported [0] [\*T\*-1] .*

ARG1: [\*-2]

ARGM-TMP: once

REL: used

ARG2: [\*] to make Kent cigarette filters

*Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.*

ARGM-TMP: recently

REL: diagnosed

ARG2: cancer

#### 1.4.7 Extent (EXT)

ARGM-EXTs indicate the amount of change occurring from an action, and are used mostly for the following:

1. Numerical adjuncts, ‘*raised prices **by 15 percent***’
2. Quantifiers such as *a lot*, *least*, *incredibly*, *extremely*, or *really*
3. Comparatives, ‘*he raised prices **more than she did***’

*PS of New Hampshire shares closed yesterday at \$ 3.75 [\*U\*], off 25 cents, in New York Stock Exchange composite trading.*

ARG1: PS of New Hampshire shares

REL: closed

ARGM-TMP: yesterday

ARGM-EXT: at \$ 3.75 [\*U\*], off 25 cents,

ARGM-LOC: in New York Stock Exchange composite trading

*‘An active 55-year-old in Boca Raton may care more about Senior Olympic games, while a 75-year-old in Panama City may care more about a seminar on health,’ she says [\*T\*-1].*

ARG0: An active 55-year-old in Boca Raton

ARGM-MOD: may

REL: care

ARGM-EXT: more

ARG1: about Senior Olympic games

ARGM-ADV: while a 75 year old in Panama City may care more about a seminar on health

*Rep. Jerry Lewis , a conservative Californian , added a provision of his own, intended [\*] to assist Bolivia, and the Senate then broadened the list further by [\*-1] including all countries in the U.S. Caribbean Basin initiate as well as the Philippines - [\*-1] backed [\*] by the powerful Hawaii Democrat Sen. Daniel Inouye.*

ARG0: the Senate

ARGM-TMP: then

REL: broadened

ARG1: the list

ARGM-EXT: further

ARGM-MNR: by [\*-1] including all countries in the U.S. Caribbean Basin initiate as well as the Philippines

ARGM-PRD: [\*-1] backed [\*] by the powerful Hawaii Democrat Sen. Daniel Inouye

Extent is also used to show the degree to which something occurs. This form of extent is most often seen in adjective annotation.

*On the other hand, the standard of living of other countries’ peoples is rising, their education levels will rise, and anti-American passion also will not be so intense.*

ARG1: anti-American passion

ARGM-ADV: also

ARGM-MOD: will

ARGM-NEG: not

ARGM-EXT: so

REL: intense

*Cdog is sometimes a bit smarter and the fact he ’s sunk to this type of argument shows what poor pickings he ’s got to work with.*

ARG1: Cdog

ARGM-TMP: sometimes

ARGM-EXT: a bit

REL: smarter

#### 1.4.8 Reciprocals (REC)

These include reflexives and reciprocals such as *himself, itself, themselves, each other, or own*, which refer back to one of the other arguments. Often, these arguments serve as the ARG1 of



the relation. In these cases, the argument should be annotated as the numbered argument as opposed to the reciprocal modifier.

*But voters decided that if the stadium was such a good idea someone would build it himself, and rejected it 59% to 41% [\*U\*].*

ARGM-ADV: if the stadium was such a good idea

ARG0: someone

ARGM-MOD: would

REL: build

ARG1: it

ARGM-REC: himself

### 1.4.9 Secondary Predication (PRD)

These are used to show that an adjunct of a predicate is in itself capable of carrying some predicate structure.

Typical examples include:

1. Resultatives, ‘*The boys pinched them **dead***’ or ‘*She kicked [the locker lid]-1 [\*-1] **shut***’
2. Depictives, ‘***Rosy-cheeked**, Santa came down the chimney*’
3. As-phrases, ‘*supplied **as security in the transaction***’

In each of these cases, it is notable that the argument labeled PRD modifies another argument of the verb (describing its state during or after the event) more than it modifies the verb or event itself.

*Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov. 29 .*

ARG0: Pierre Vinken , 61 years old ,

ARGM-MOD: will

REL: join

ARG1: the board

ARGM-PRD: as a nonexecutive director

ARGM-TMP: Nov. 29

*Prior to his term , a teacher bled to death in the halls , [\*-1] stabbed [\*-2] by a student.*

ARGM-TMP: Prior to his term

ARG1: a teacher

REL: bled

ARGM-PRD: to death

ARGM-LOC: in the halls

ARGM-ADV: [\*-1] stabbed [\*-2] by a student

*This wage inflation is bleeding the NFL dry, the owners contend [\*T\*-1].*

ARG0: This wage inflation

REL: bleeding

ARG1: the NFL

ARGM-PRD: dry

### 1.4.10 Purpose Clauses (PRP)

Purpose clauses are used to show the motivation for some action. Clauses beginning with ‘*in order to*’ and ‘*so that*’ are canonical purpose clauses.

*More than a few CEOs say [0] the red-carpet treatment tempts them to return to a heartland city for future meetings.*

ARG1: them

REL: return

ARG4: to a heartland city

ARGM-PRP: for future meetings

*In a disputed 1985 ruling , the Commerce Commission said [0] Commonwealth Edison could raise its electricity rates by \$ 49 million [\*U\*] [\*-1] to pay for the plant.*

ARG0: Commonwealth Edison

ARGM-MOD: could

REL: raise

ARG1: its electricity rates

ARG2: by \$ 49 million [\*U\*]

ARGM-PRP: [\*-1] to pay for the plant

#### 1.4.11 Cause Clauses (CAU)

Similar to purpose clauses, these indicate the reason for an action. Clauses beginning with ‘because’ or ‘due to’ are canonical cause clauses. Questions starting with ‘Why,’ which are always characterized by a trace linking back to this question word, are always treated as cause. However, in these question phrases it can often be difficult or impossible to determine if the ‘why’ truly represents purpose or cause. Thus, as a general rule, if the annotator cannot determine whether an argument is more appropriately purpose or cause, cause is the default choice.

*Pro-forma balance sheets clearly show why-1 Cray Research favored the spinoff [\*T\*-1] .*

ARGM-CAU: [\*T\*-1]

ARG0: Cray Research

REL: favored

ARG1: the spinoff

*However , five other countries – China , Thailand , India , Brazil and Mexico – will remain on that so-called priority watch list because of an interim review , U.S. Trade Representative Carla Hills announced [0] [\*T\*-1] .*

ARGM-DIS: However

ARG1: five other countries – China , Thailand , India , Brazil and Mexico –

ARGM-MOD: will

REL: remain

ARG3: on that so-called priority watch list

ARGM-CAU: because of an interim review

#### 1.4.12 Discourse (DIS)

These are markers which connect a sentence to a preceding sentence. Examples of discourse markers are: *also, however, too, as well, but, and, as we’ve seen before, instead, on the other hand, for instance*, etc. Additionally, vocatives wherein a name is spoken (e.g. ‘**Alan**, will you go to the store?’) and interjections (e.g. ‘**Gosh**, I can’t believe it’) are treated as discourse modifiers. Because discourse markers add little or no semantic value to the phrase, a good rule of thumb for deciding if an element is a discourse marker is to think of the sentence without the potential discourse marker. If the meaning of the utterance is unchanged, it is likely that the element can be tagged as discourse.

Note that conjunctions such as *but, or*, and *and* are only marked in the beginning of the sentence. Additionally, items that relate the instance undergoing annotation to a previous sentence such

as *however*, *on the other hand*, *also*, and *in addition*, should be tagged as ARGM-DIS. However, these elements can alternately be tagged as ARGM-ADV when they relate arguments within the clause being annotated (e.g. ‘*Mary reads novels in addition to writing poetry*’) as opposed to relating to or juxtaposing an element within the sentence to an element outside the sentence (e.g. ‘*In addition, Mary reads novels*’). Often, but not always, when these elements connect the annotation instance to a previous sentence, they occur at the beginning of the instance. Consider these examples to clarify this difference:

*But for now , they ’re looking forward to their winter meeting – Boca in February.*

ARGM-DIS: But

ARGM-TMP: for now

ARG0: they

REL: [looking] [forward]

ARG1: to their winter meeting – Boca in February

*The notification also clarifies the requirements of the evaluation.*

ARG0: The notification

ARGM-DIS: also

REL: clarifies ARG1: the requirements of the evaluation.

*The notification recognizes the company and also clarifies the requirements of the evaluation.*

ARG0: The notification

ARGM-ADV: also

REL: clarifies

ARG1: the requirements of the evaluation.

Remember, do not mark *and*, *or*, or *but* when they connect two clauses in the same sentence.

As previously mentioned, another type of discourse marker includes vocatives, which are marked as VOC in TreeBank:

*TreeBank annotation:*

```
(S (NP-VOC Kris),
  (NP-SBJ *)
  (VP go
    (ADVP-DIR home)))
```

*PropBank annotation:*

ARGM-DIS: Kris

REL: go

ARG0: [\*]

ARGM-DIR: home

Vocative NPs in imperative sentences as shown above should not be tagged as coreference chains (e.g. ARG0: [\*] \* [Kris]) in order to make annotation consistent with other examples of vocative NPs, which do not include traces:

*I ain't kidding you, Vince*

ARGM-DIS: Vince

REL: kidding

ARG0: I

ARG1: you

ARGM-NEG: n't

And, finally, the class of discourse markers includes interjections such as *oh my god*, *ah*, and *damn*.

*I might point out that your inability to report to my office this morning has not ah limited my knowledge of your activities as you may have hoped.*

ARGM-DIS: ah

REL: limited

ARGM-NEG: not

ARG1: my knowledge of your activities

ARG0: your inability to report to my office this morning

ARGM-ADV: as you may have hoped

### 1.4.13 Modals (MOD)

Modals are: *will*, *may*, *can*, *must*, *shall*, *might*, *should*, *could*, and *would*. These elements are consistently labeled in the TreeBank as MOD. These are one of the few elements that are selected and tagged directly on the modal word itself, as opposed to selecting a higher node that contains the lexical item.

### 1.4.14 Negation (NEG)

This tag is used for elements such as *not*, *n't*, *never*, *no longer*, and other markers of negative sentences. Negation is an important notion for PropBank annotation; therefore, all markers which indicate negation should be marked as ARGM-NEG. For example, when annotating adverbials like *never*, which could be marked as either TMP or NEG, the NEG tag should be used. These are also elements that are tagged directly on the lexical item itself as opposed to on a higher node. Be careful to distinguish these from the conjunction *not only*, which does not actually indicate that the verb is negative and should not be annotated because it is a conjunction.

### 1.4.15 Direct Speech (DSP)

A verb of expression is any verb which has a speaker/thinker argument (ARG0) and the utterance/thought (ARG1). If the utterance is a constituent, then there is a trace in TreeBank which is coindexed with that constituent. PropBank annotation tags the trace as ARG1 in this case:

*TreeBank Annotation:*

```
(S ‘‘
  (S-TPC-1 (NP-SBJ We)
    (VP will
      (VP win)))
  ,
  ,
  (NP-SBJ Mary)
  (VP said
    (S *T*-1))
  .))
```

*PropBank Annotation:*

REL: said

ARG1: [\*T\*-1]

ARG0: Mary

Unfortunately, in many examples, the utterance does not correspond to one constituent in TreeBank:

*Among other things , they said [??] , Mr. Azoff would develop musical acts for a new record label.*

*TreeBank Annotation:*

```
(S
  (PP (IN Among)
    (NP (JJ other) (NNS things) ))
  (PRN
    ( , , )
    (S
      (NP-SBJ (PRP they) )
      (VP (VBD said)
        (SBAR (-NONE- 0)
          (S (-NONE- ***) ))))
    ( , , ) )
  (NP-SBJ (NNP Mr.) (NNP Azoff) )
  (VP (MD would)
    (VP (VB develop)
      (NP
        (NP (JJ musical) (NNS acts) )
        (PP (IN for)
          (NP (DT a) (JJ new) (NN record) (NN label) )))))
  ( . . ) )
```

As the example above shows, in such cases, the ARG1 argument of the verb say is a \*\*\* empty category, which does not have an index in TreeBank. PropBank annotation tags this empty category as ARG1 in this case; however, it also provides a link between this empty category and the top S node, which contains the utterance as well as the verb of saying. This is a rare exception wherein the annotation will include embedded arguments. Thus, the annotator first selects and tags the entire SBAR node as ARG1. Next, the annotator selects the empty trace itself NONE-\*\*\* and annotates this node as ARGM-DSP. While this annotation is still in the Jubilee memory, the annotator subsequently selects the S-node containing the *rel* and uses the \* link to link the two together: click **Argument** on the Jubilee menu bar followed by clicking **Functions**. From the options therein, select \* (shortcut: **Ctrl+Shift+8**). Although seasoned annotators should feel uncomfortable creating embedded, recursive annotations such as this, there is special post-processing for these cases that effectively removes what is often a PRN (parenthetical) node containing the relation and its arguments from the argument that is semantically ‘*what is spoken.*’ In the final version of the PropBank, ARGM-DSP tag will be replaced by LINK-DSP, to indicate that this is not a modifier of the verb, but simply additional information about one of its arguments. Figure 1.3 shows the correct annotation of an instance of DSP as it will be seen in Jubilee.

ARG1: [SBAR (-NONE- 0) S (-NONE- \*\*\*)]

ARG0: they

ARGM-DSP: [-NONE-\*\*\*] \* [Among other things , they said [??] , Mr. Azoff would develop musical acts for a new record label]

REL: said

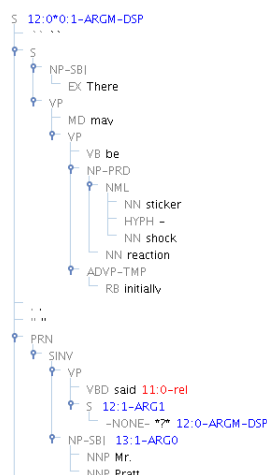


Figure 1.3: Correct annotation of DSP

### 1.4.16 Adverbials (ADV)

These are used for syntactic elements which clearly modify the event structure of the verb or adjective in question, but which do not fall under any of the headings above. The annotator should always try to use one of the alternate modifiers listed above before assuming that a modifier is merely adverbial. However, adverbial elements are often...

1. Temporally related (modifiers of events): *Treasures are just lying around, **waiting to be picked up***
2. Intensional (modifiers of propositions): *probably, possibly*
3. Focus-sensitive *only, even*
4. Sentential (evaluative, attitudinal, viewpoint, performatives): *fortunately, really, legally, frankly speaking*, clauses beginning with *given that, despite, except for, or if*

As opposed to ARGM-MNR, which modify the verb, ARGM-ADVs usually modify the entire sentence. In some cases, modifiers like *happily* can be ambiguous between MNR and ADV interpretations, as shown below:

*She sang happily.*

ARGM-MNR: happily

*Happily, she sang.* (paraphrase: ‘*I am happy that she sang*’)

ARGM-ADV: happily

In these cases, use context as much as possible to try to make the best judgment.

### 1.4.17 Adjectival (ADJ)

This tag is used in a manner that is similar to ADV, but its use is restricted to noun annotation. See Chapter 3 and Section 3.3.1 for more details.

### 1.4.18 Light Verb (LVB)

This tag is used to label the light verb only in the noun pass of light verb annotation. See Chapter 2 for more details.

### 1.4.19 Construction (CXN)

This tag is used to label arguments that are projected by a construction, such as the comparative construction (e.g. *She is taller than her sister*), as opposed to a lexical relation. Although constructions can project arguments with a variety of lexical relations, this tag is currently only used in the annotation of adjective relations. See Chapter 4 for more details.

## 1.5 Span of Annotation

For the purposes of PropBank annotation, annotators should only assign arguments within a certain syntactic span surrounding the **rel**. The structure of the tree reflects which constituents in an utterance are truly arguments of a particular predicate; thus, even when annotators feel that a constituent outside of this span has some semantic bearing on the **rel**, it should not be annotated. Rather, the syntactic span of annotation should be respected: everything within that span should be encompassed by an argument label (with exceptions described below), and nothing outside of that span should be annotated (with exception of linking annotation, such as that of relative clauses).

Do not tag determiners (labeled **DT** in TreeBank) or conjunctions (labeled **CC** or **CONJP** in TreeBank), unless these begin the sentence and are being used in a discourse function, as described in Section 1.4.12. Do not tag auxiliary verbs such as *have*, *be*, or *do*; the auxiliary verb itself will come up for annotation and at that point the auxiliary sense will be selected without further annotation.

Tag all and only the following:

1. Sisters of the verb
2. Sisters of the VP

To determine the span of annotation, locate the **rel** and the accompanying TreeBank tag indicating one of the following types of verbs <sup>1</sup>:

VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VCN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Figure 1.4 shows the TreeBank view of a typical instance in Jubilee, with the **VB** node indicated.

Once this has been located, annotate the sisters of this node. Sisters to the verb will be parallel to it in the tree. Figure 1.5 has an arrow where the annotator should look, beginning at the verb, for sisters. In this case, it is a relatively short distance and only the **NP-EXT** node needs to be annotated as **ARG1**; however, in some cases, the annotator will have to scroll up and down through Jubilee's TreeBank view to annotate multiple sisters to the verb.

Next, examine the tree to see if the **VB** node is embedded in a **VP** (verb phrase) node. The verb is usually located inside a higher **VP** node, unless it is located inside an **NP** (noun phrase) node. Where the verb is accompanied by one or more auxiliaries, it may be encompassed by several **VP** nodes, as illustrated by Figure 1.6, which indicates each of the **VP** nodes:

<sup>1</sup>For a complete listing of TreeBank tags, see [http://www.ling.upenn.edu/courses/Fall\\_2003/ling001/penn\\_treebank\\_pos.html](http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

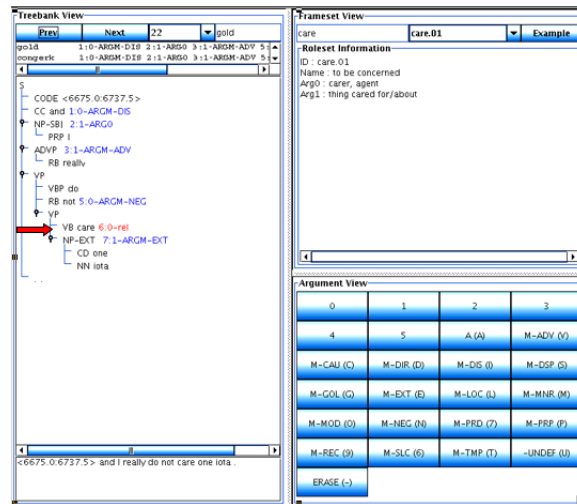


Figure 1.4: Find the rel

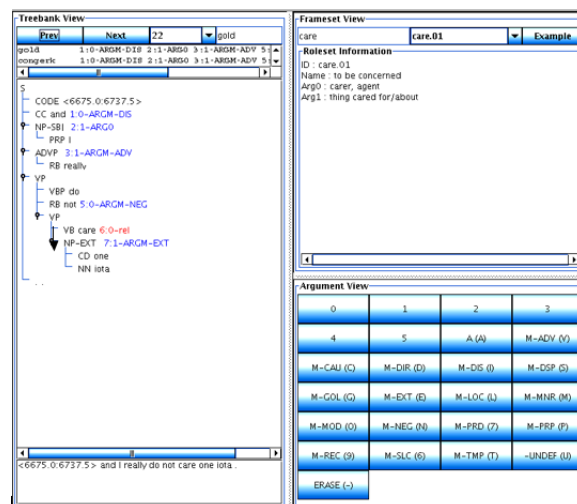


Figure 1.5: Find and annotate any sisters to the rel

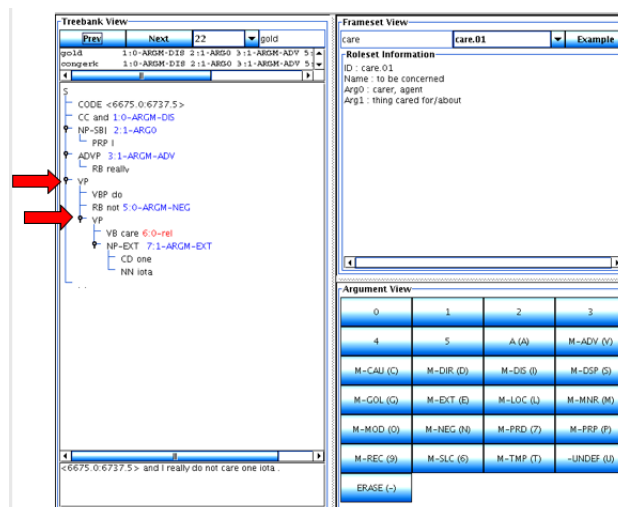


Figure 1.6: Find the highest VP node containing the rel



Once the annotator has located the highest VP node, annotate all sisters to each VP node (again, the nodes that are parallel to VP nodes in the tree). Figure 1.7 uses arrows to illustrate where the annotator should look, beginning at the verb, for sisters to the VP nodes:

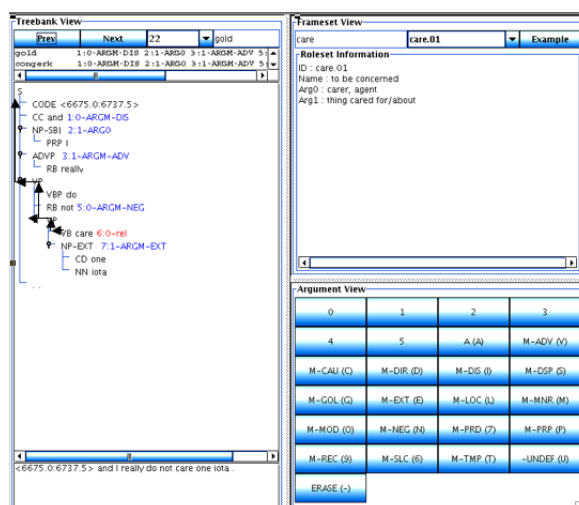


Figure 1.7: Find and annotate any sisters to the VP nodes containing the **rel**

Thus, following this line, the annotator discovers several nodes that must be annotated: firstly, the RB node as ARGM-NEG; the auxiliary *do* can be ignored; second, the ADVP node as ARGM-ADV; next the NP-SBJ node as ARG0; finally the CC node as ARGM-DIS because it meets the conditions described in Section 1.4.12 for serving a discourse function. The CODE node can be ignored when present. Additionally, TOP nodes at the very top of an instance should never be annotated.

The last thing to note is that when the verb is embedded in a VP node, '*S marks the spot*' to stop annotation:

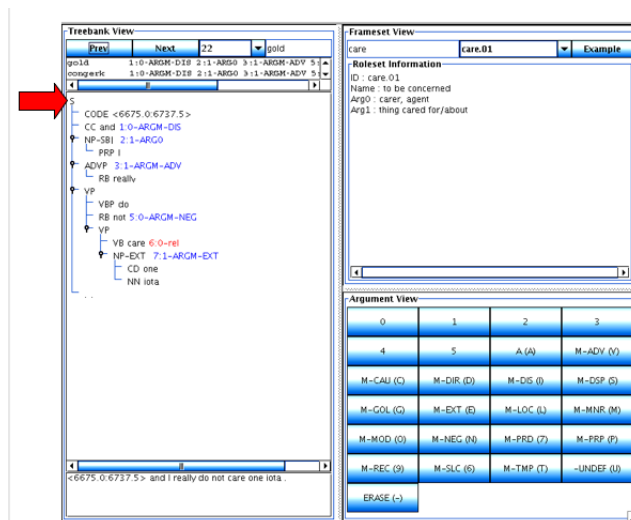


Figure 1.8: Stop at S

S indicates clausal boundaries in the TreeBank. Thus, anything beyond the S would also be beyond the clause containing the **rel**, and in turn, constituents outside of this clause are not arguments of the **rel**. Only linking practices should require attention to constituents outside of the S node containing the **rel**. TreeBank annotation can mark clausal boundaries with any

of the following tags, indicating what type of clausal boundary it is; all should be treated as marking one end of the annotation span:

- 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**.

Note that **S** nodes can also serve as sentential complements to a verb, as seen in Figure 1.9. Which **S** node is of focus when determining the span depends on the relative location of the **rel**.

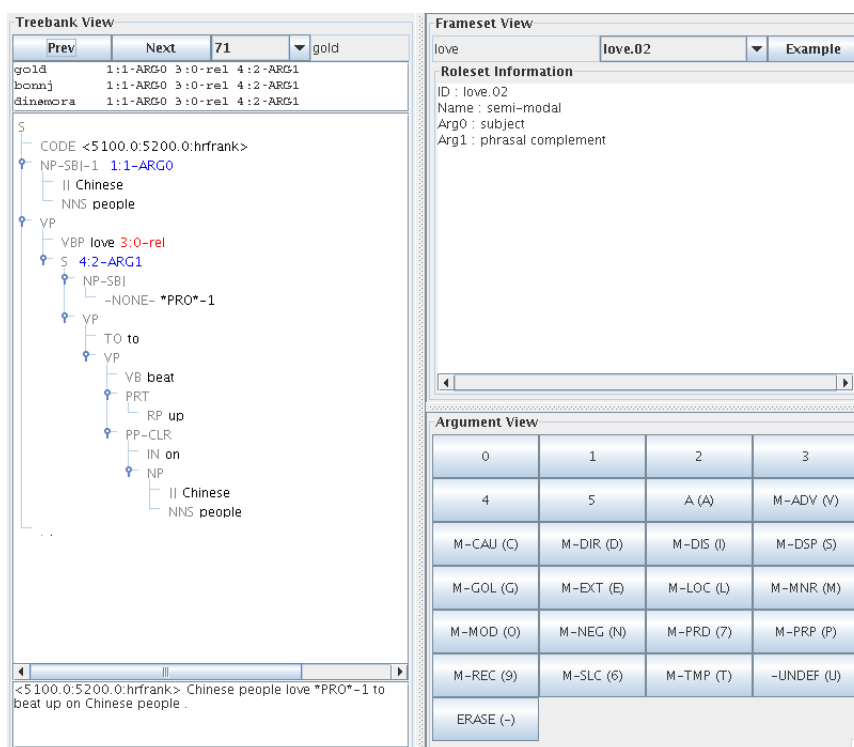


Figure 1.9: **S** as sentential complement

As mentioned previously, the **rel** can potentially arise in other types of nodes, such as NP (noun phrase) or ADJP (adjective phrase) nodes, with or without an intervening VP node. When this happens, the annotation span cannot be thought of as delimited by an **S** node. Just as in previously mentioned cases, annotate sisters to the verb or VB TreeBank tag and annotate sisters to the VP node when present. Although the **S** node will not be present as a clue for where to stop annotation, simply do not annotate constituents that are parents or aunts to the verb or verb phrase node. Parents and aunts, unlike sisters, will not be along a parallel line with the verb or verb phrase. Instead, their root nodes will be located to the left of this line. This heuristic applies when determining the span of any annotation, but may be especially important in the absence of an **S** node.

Figure 1.10 gives an example of a **rel** contained within an NP node. Note that the NP node

itself and everything outside of this node should not be annotated because these constituents are parents or aunts of the verb phrase rather than sisters of the verb phrase or verb.

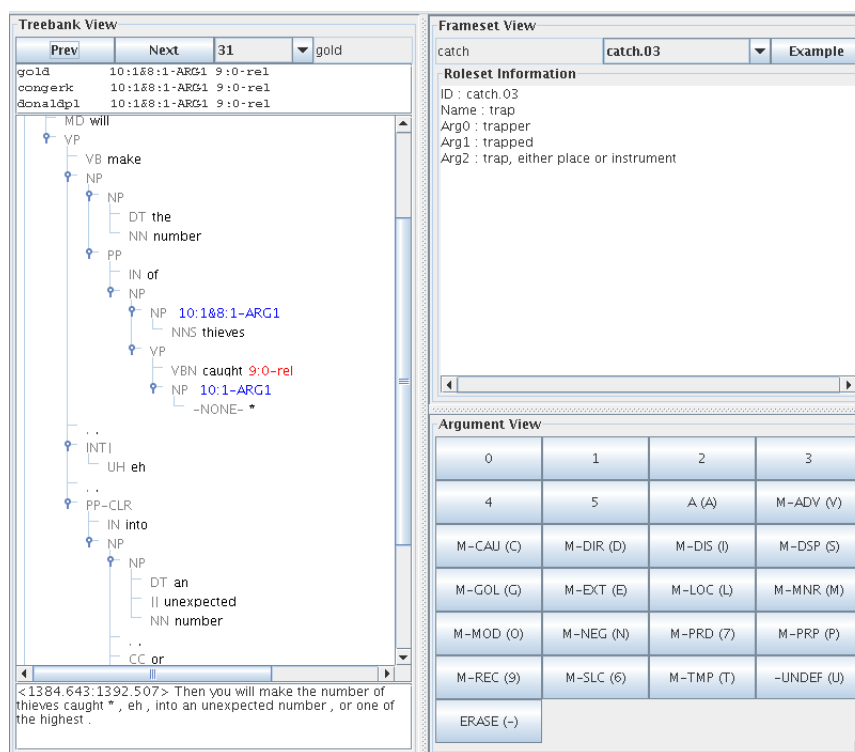


Figure 1.10: *rel* embedded in an NP node

## 1.6 Where to Place Tags

Jubilee allows you to select any node in the tree; thus, it is up to the annotator to select the appropriate node reflecting the correct constituent boundaries of an argument. In general, the node above the lexical item itself, which indicates the syntactic function of that constituent (e.g., NP, NP-SBJ, PP, ADJP, ADVP, etc.), is the correct placement for the tag. However, as mentioned in the ARGM sections, annotation of modals and negatives require placement of the tag directly on the lexical item because there is no higher node to annotate without including more than just the modal or negative marker. Please review the Jubilee screen shots given in Section 1.5 for examples of the correct placement of tags.

Occasionally, the phrase structure of an instance is such that the annotator must choose between annotating a higher node as a single argument or annotating several nodes embedded therein as various arguments. As a general rule of thumb, if it is possible to place a lower-numbered argument tag on a single, higher node, this is preferable to annotating several higher-numbered arguments and/or modifiers on embedded nodes therein. For example, Figure 1.11 below is correct given the argument structure outlined in the frame *hold.04*, but it is dispreferred to Figure 1.12, which simply tags the higher node with a lower argument number.

### 1.6.1 Exceptions to Normal Tag Placement

Certain verbs such as *encourage* and *persuade*, which involve both an impelled agent and an impelled action require that annotators break up and delve into the sister *S* node in order to annotate the impelled agent and impelled action separately. These are often cases of verbs

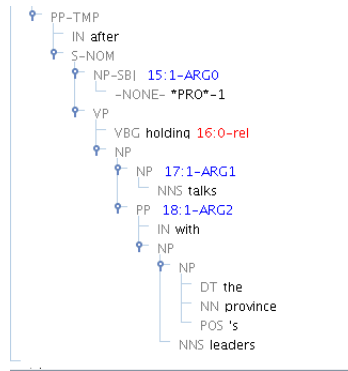


Figure 1.11: Theoretically correct, but dispreferred annotation

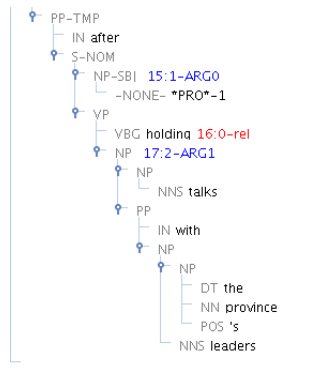


Figure 1.12: Preferred annotation of higher node with lower argument number

that participate in exceptional case marking, meaning that the matrix verb (e.g., *encourage* or *persuade*) assigns the accusative case to the subject of the sentential complement:

*I encouraged **him** to annotate for PropBank.*

*\*I encouraged **he** to annotate for Propbank.*

Because the matrix verb assigns accusative case to the subject of the infinitival complement, it is thought that the verb assigns semantic roles to both the agent and action separately. In some cases, such usages are parsed with two separate constituents that can be marked as usual with separate arguments: an NP node for the impelled agent and an S node for the action. However, in other cases, the sentential complement of these verbs forms a single constituent, which is an S node that is a sister to the matrix verb. As a result, where we would normally only annotate the S node with an argument, in these cases, we annotate within the S node to separately tag the impelled agent and impelled action. These are essentially theoretical disagreements in how to parse such instances. As always, annotators should allow the rolesets to guide their annotations.

## 1.7 Understanding and Annotating Null Elements in the Penn TreeBank

The inventory of null elements used in Penn TreeBank is as follows:

[*PRO*]	overt subjects, subject control, and small clauses
[*]	passive traces including reduced relative clauses and raising constructions
[*T*]	trace of A-movement, including parasitic gaps
[(NP *)]	arbitrary PRO, controlled PRO, and trace of A-movement
[0]	null complementizer, including null wh-operator
[*U*]	unit
[*?*]	placeholder for ellipsed material
[*NOT*]	anti-placeholder in template gapping
[*RNR*]	pseudo-attach: right node raising
[*ICH*]	pseudo-attach: interpret constituent here
[*EXP*]	pseudo-attach: extraposition

This section presents some examples of most commonly used null elements and their PropBank annotation.

### 1.7.1 Passive Sentences

Sentences can be either active (*The executive committee approved the new policy*) or passive (*The new policy was approved by the executive committee*). In active sentences, the subject is the agent or do-er of the action, marked as ARG0 in PropBank. In passive sentences, the subject of the sentence is acted upon by some other agent or by something unnamed, and is being marked as ARG1 in PropBank.

Passive sentences are assumed to be derived from the corresponding active sentences by movement of the object to the subject position. This movement leaves a trace, represented as [\*] in TreeBank. Except in the case of reduced relatives, this trace will already be coindexed with its realized referent:

Active: *Mary hit John*

Passive: *John-1 was hit [\*-1] by Mary.*

Since TreeBank provides a link between [\*-1] and *John*, it is the trace, rather than the NP *John*, which is being labeled as ARG1 in PropBank:

*PropBank annotation:*

REL: hit

ARG1: [\*-1]

ARG0: by Mary

The following example illustrates a TreeBank representation of the passive sentence. The link between the trace and the NP is indicated by the number 1 in the trace (NP-3 \*-1) and (NP-SBJ-1 he) below. Note that chains of coreference are represented in the Penn TreeBank using various numerals, and that one element such as ‘he’ can potentially be semantically present in several positions in the underlying syntax; therefore, several numbered indices may be connected to one element. It is important to follow the chains of coreference throughout the instance to ensure a full understanding of each null element.

*TreeBank annotation:*

```
(S (NP-SBJ-1 he)
  (VP was
    (VP accused
      (NP-3 *-1)
      (PP-CLR of
        (S-NOM (NP-SBJ *-3)
          (VP (VP conducting
```

```

                (NP illegal business))
and
(VP possessing
  (NP illegal materials)))))))))

```

Again, it is the trace which is being annotated as the argument:

*PropBank annotation:*

ARG1: [NP-3 \*-1]

REL: accused

ARG2: of [\*3\*] conducting illegal business and possessing illegal materials

### 1.7.2 Fronted and Dislocated Arguments

Other examples of moved constituents are fronted or otherwise dislocated arguments and adjuncts. As in the other cases of movement, fronted elements leave a trace, which is being coindexed with the moved constituent in TreeBank.

In the following example, the ARG2 (*‘where put’*) argument of the verb *put* is being fronted. In the TreeBank annotation, this is indicated by the chain which links the trace [\*T\*-1] with the adverbial *There*:

*TreeBank annotation:*

```

(S (ADVP-PUT-TPC-1 There)
  ,
  (NP-SBJ I)
  (VP put
    (NP the book)
    (ADVP-PUT *T*-1) ))

```

As with annotation of passive traces, the ARG2 argument is the trace, rather than the fronted constituent:

*PropBank annotation:* REL: put

ARG0: I

ARG1: the book

ARG2: [\*T\*-1]

Modifiers, or ARGMs, can be fronted as well, as the following example shows:

*TreeBank annotation:*

```

(S (SBAR-PRP-TPC-9 Because
  (S (NP-SBJ I)
    (VP 'm
      (NP-PRD such a bad boy))))
  (NP-SBJ I)
  (VP think
    (SBAR 0
      (S (NP-SBJ I)
        (VP wo n't
          (VP get
            (NP a lollipop)
            (SBAR-PRP *T*-9) )))))

```

Since the ‘*because*’ clause modifies the verb *get* in this example, the trace originates as the modifier of *get*. This trace is being annotated as ARGM-CAU in PropBank:

*PropBank annotation:*

REL: get  
 ARG1: a lollipop  
 ARG0: I  
 ARGM-NEG: n’t  
 ARGM-MOD: wo  
 ARGM-CAU: [\*T\*-9]

In rare situations, movement does not leave a trace, but rather leaves a pronoun (called a resumptive pronoun). In such cases, the argument of the verb is a higher NP, which includes both the pronoun and the trace to the topicalized NP in TreeBank. This NP is annotated as ARG1 in PropBank:

*TreeBank annotation:*

```
(S (NP-TPC-1 John)
    ,
    (NP-SBJ I)
    (VP like
      (NP (NP him)
          (NP-1 *T*))
      (NP-ADV a lot))))
```

*PropBank annotation:*

REL: like  
 ARG0: I  
 ARG1: [NP (NP him) (NP-1 \*T\*)]  
 ARGM-MNR: a lot

In even more rare situations, the topicalized NP and the pronoun are not already co-indexed in the TreeBank. See Section 1.8.4 for further description of how to annotate these instances.

### 1.7.3 Questions and Wh-Phrases

Another type of traces is a trace of a wh-phrase in questions.

*What do you like?*

As in the case of passive sentences, questions are assumed to be derived by movement. In the example below, the ARG1 argument of the verb *like* is a wh-phrase *what*, which moves from the object position of the verb to the front of the sentence. This movement leaves a trace, as shown below:

*What-1 do you like [\*T\*-1]?*

In TreeBank annotations, wh-phrases are marked as WHNP. As in the case of passive sentences, TreeBank provides a link between the trace and the moved WHNP:

*TreeBank annotation:*

```
(SBARQ (WHNP-1 what)
  (SQ do
    (NP-SBJ you)
    (VP like
      (NP *T*-1))))
```

Again, for the purposes of PropBank, the argument ARG1 is the trace, as shown below:

*PropBank annotation:*

REL: like

ARG0: you

ARG1: [\*T\*-1]

Wh-phrases are not necessarily core arguments. However, questions can be formed with wh-phrases like *when*, *where*, or *how*, in which case they should be tagged as ARGMs.

*TreeBank annotation:*

```
SBARQ (WHNP-1 Which day)
      (SQ did
        (NP-SBJ you)
        (VP get
          (ADVP-DIR there)
          (NP-TMP *T*-1)))
```

*PropBank annotation:*

ARG0: you

REL: get

ARG2: there

ARGM-TMP: [\*T\*-1]

*TreeBank annotation:*

```
(SBARQ (WHADVP-42 How)
      (SQ did
        (NP-SBJ you)
        (VP fix
          (NP the car)
          (ADVP-MNR *T*-42)))
      ?)
```

*PropBank annotation:*

REL: fix

ARG0: you

ARG1: the car

ARGM-MNR: [\*T\*-42]

Questions can also be embedded, as in the example below. PropBank annotation is not different from direct questions in this case:

*John didn't know where-3 his parents had met [\*T\*-3].*

ARG0: his parents

REL: met

ARGM-LOC: [\*T\*-3]

#### 1.7.4 Interpret Constituent Here (ICH) Traces

ICH traces are used in TreeBank to indicate a relationship of constituency between elements separated by intervening material. An example of such split constituents are ‘heavy shift’ constructions, illustrated below:

*TreeBank annotation:*



```

(S (NP-SBJ (NP a young woman)
            (SBAR *ICH*-1))
  (VP entered
    (SBAR-1 (WHNP-2 whom)
      (S (NP-SBJ she)
        (PP-TMP at
          (ADVP once))
        (VP recognized
          (NP *T*-2)
          (PP-CLR as
            (NP Jemima Broadwood))))))))))

```

The subject NP in this case is being split into two constituents: the NP ‘*a young woman*’ and SBAR ‘*whom she at once recognized as Jemima Broadwood.*’ The ICH trace specifies a link to the SBAR node in this example. Essentially, the NP in addition to the material linked by ICH trace can be thought of as one whole constituent: ‘*A young woman whom she at once recognized as Jemima Broadwood,*’ part of which has been moved for pragmatic purposes.

In all examples of this type, the argument is the constituent which includes the ICH trace:

*PropBank annotation:*

ARG0: a young woman [ICH-1]

REL: entered

It is very important that the annotator does not annotate the dislocated part of the constituent (in the previous case the SBAR-1 material) a second time with another tag. Underlyingly, the material connected by ICH trace is part of the NP ‘*a young woman,*’ which is already annotated as ARG0. In post-processing, the rest of this constituent linked by ICH trace will be concatenated to the NP annotated as ARG0; thus, tagging the dislocated portion of the constituent a second time will create recursive annotation and will be returned as an error. In other words, tagging the dislocated portion will be recognized computationally as having a second argument of a different type embedded in the first, which is disallowed.

Other typical examples of ICH traces are shown below:

*[Five ICH-1] ran, [out of the twenty-five that showed up]-1.*

ARG0: Five \*ICH-1\*

REL: ran

*[Some people in Paris]-1 want \*PRO\*-1 to hear more [ICH-2] from me [than those fellers over at the conference house do]-2.*

ARG0: \*PRO\*-1

REL: hear

ARG1: more [ICH-2]

ARG2: from me

Figure 1.13 shows correct annotation for an instance containing an ICH trace.

### 1.7.5 Right Node Raising (RNR) Traces

RNR traces are used when a constituent is interpreted simultaneously in more than one place. An example of a right node raising structure is given below:

*TreeBank annotation:*

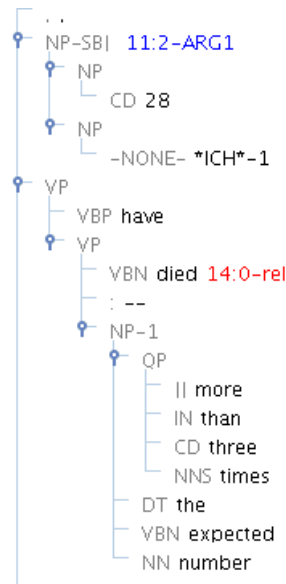


Figure 1.13: Correct annotation of ICH

```
(NP (NP (ADJP so many) enchained demons)
  (VP straining
    (PP-MNR in
      (NP anger))
    (S (NP-SBJ *)
      (VP to
        (VP (VP tear
              (NP *RNR*-1))
          and
          (VP gnaw
            (PP-CLR on
              (NP *RNR*-1)))
            (NP-1 his bones))))))
```

In this example, the NP ‘his bones’ is interpreted as both the argument of the verb *tear* and the verb *gnaw*. When annotating the verb *tear*, the trace, (NP \*RNR\*-1), is the argument of the verb:

*PropBank annotation:*

REL: tear

ARG1: [\*RNR\*-1]

ARG0: [NP-SBJ\*]

Likewise, when annotating the verb *gnaw*, the prepositional phrase, including the trace (PP-CLR on (NP \*RNR\*-1))), is analyzed as the argument:

*PropBank annotation:*

REL: gnaw

ARG1: on [\*RNR\*-1]

ARG0: [NP-SBJ\*]

A similar annotation applies when the RNR trace is a clausal argument:

*I want \*RNR\*-1 and like \*RNR\*-1 [\* to eat ice-cream]-1.*

ARG0: I  
REL: want  
ARG1: \*RNR\*-1

If the RNR trace is part of the argument of the verb, then select the argument including the trace:

*His dreams had revolved around her so much and for so long that...*

*TreeBank annotation:*

```
(S (NP-SBJ His dreams)
  (VP had
    (VP revolved
      (PP-CLR around
        (NP her))
      (UCP-ADV (ADVP (ADVP so much)
        (SBAR *RNR*-1))
        and
        (PP-TMP for
          (NP (NP so long)
            (SBAR *RNR*-1)))
        (SBAR-1 that...))))))
```

*PropBank annotation:*

ARG1: his dreams  
REL: revolved  
ARGM-LOC: around her  
ARGM-EXT: so much [\*RNR\*]

The following example illustrates annotation of RNR traces within a small clause (for further information on the annotation of small clauses, see Section 1.9).

*But our outlook has been and continues to be defensive*

*TreeBank annotation:*

```
(S But
  (NP-SBJ-2 our outlook)
  (VP (VP has
    (VP been
      (ADJP-PRD *RNR*-1)))
    ,
    and
    (VP continues
      (S (NP-SBJ *-2)
        (VP to
          (VP be
            (ADJP-PRD *RNR*-1))))
      ,
      (ADJP-PRD-1 defensive))))
```

*PropBank annotation:*

REL: continue  
ARG1: [\*-2] to be \*RNR-1

Figure 1.14 shows correct annotation of an instance containing RNR traces.

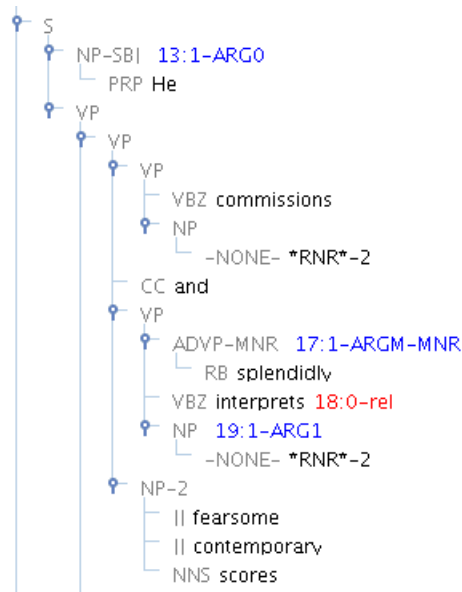


Figure 1.14: Correct annotation of RNR

### 1.7.6 It EXtraPosition (EXP)

Dummy placeholders in English such as *it* or *that* do not add any meaning to the sentence. In the following example, the syntactic subject of the sentence is a dummy *it*, which includes a trace *\*EXP\*-1*. This trace refers to the logical, semantic subject of the sentence, marked as SBAR-1:

*TreeBank annotation:*

```
(S (NP-SBJ (NP It)
           (SBAR *EXP*-1))
   (VP is
      (ADJP-PRD clear)
      (PP to
         (NP me))
      (SBAR-1 that
         (S (NP-SBJ this message)
            (VP is
               (ADJP-PRD unclear))))))
```

In PropBank annotations, dummy *it* and EXP traces are **not included, do not tag them**:

*PropBank annotation:*

REL: is  
 ARG1: that this message is unclear  
 ARG2: clear to me

Rather, tag only that which has semantic value in the utterance, the overt constituent. Thus, the underlying phrase can be thought of semantically as: ‘*That this message is unclear is clear to me.*’ The *it* is merely added for pragmatic purposes to avoid having such a heavy constituent at the front of the phrase.

Another example:

*It required an energy he no longer possessed to be satirical about his father.*

*PropBank annotation:*

ARG0: to be satirical about his father

ARG1: an energy he no longer possessed

REL: required

In the examples below, the dummy constituents are the objects, rather than the subjects. As in the case of dummy subjects, only the logical argument is being tagged, whereas the dummy pronoun and the EXP trace are not part of the PropBank annotation:

*Mrs. Yeargin was fired [\*-1] and prosecuted [\*-1] under an unusual South Carolina law that-79 [\*T\*-79] makes it [\*EXP\*-2] a crime [\*] to breach test security.*

*PropBank annotation:*

ARG0: [\*T\*-79]

REL: makes

ARG2: a crime

ARG1: [\*] to breach test security

*Any raider would find it [\*EXP\*-1] hard [\*] to crack AG 's battlements.*

*TreeBank annotation:*

```
(S
  (NP-SBJ (DT Any) (NN raider) )
  (VP (MD would)
    (VP (VB find)
      (S
        (NP-SBJ
          (NP (PRP it) )
          (S (-NONE- *EXP*-1) ))
        (ADJP-PRD (JJ hard) )
        (S-1
          (NP-SBJ (-NONE- *) )
          (VP (TO to)
            (VP (VB crack)
              (NP
                (NP (NNP AG) (POS 's) )
                (NNS battlements) ))))))))
  (. .) )
```

*PropBank annotation:*

ARG0: Any raider

ARGM-MOD: would

REL: find

ARG3: hard

ARG1: [\*] to crack AG 's battlements

Figure 1.15 shows correct annotation of an instance of the copular sense of *to be*, which contains it EXP.

Common mistake: Please make sure to distinguish dummy *it* from the referring pronoun *it*, where *it* refers to a previous NP, a clause, or an event. (hint: referring pronouns are not followed by an EXP trace in TreeBank). All referring pronouns, including *it*, should be marked as arguments in PropBank.

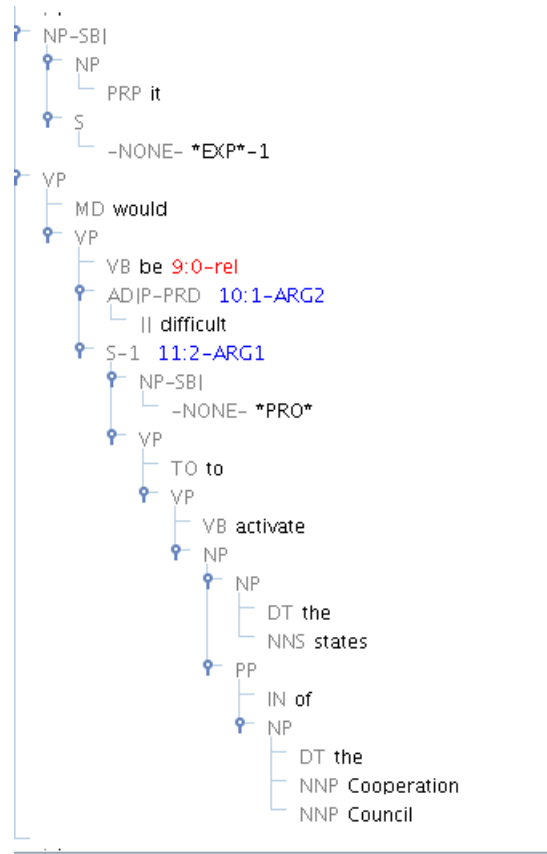


Figure 1.15: Correct annotation of EXP and be.01

*It sounds good.*

REL: sounds

ARG1: it

ARGM-MNR: good

*Italy's Foreign Ministry said [0] it is investigating exports to the Soviet Union.*

REL: investigating

ARG0: it

ARG1: exports to the Soviet Union

### 1.7.7 Other Traces

Other types of traces include the null complementizer trace, 0, the ? trace (used in ellipsis constructions), and the PPA trace in cases of predictable ambiguous attachments.

Null complementizer traces should be included as part of the clausal argument; thus, the ARG1 in this case would be annotated at the level of the SBAR node:

*TreeBank annotation:*

```
(S (NP-SBJ I)
  (VP believe
    (SBAR 0
      (S (NP-SBJ you)
        (VP are
          (ADJP-PRD smart))))))
```

*PropBank annotation:*  
REL: believe  
ARG0: I  
ARG1: [[0] you are smart]

## 1.8 Linking and Annotation of Null Elements

### 1.8.1 Relative Clause Annotation

Relative clauses are clauses that modify an N or an NP as in ‘*answers that we’d like to have.*’ Relative clauses also include a trace, which is coindexed with the relativizer in TreeBank (e.g., *that, which, who*). Alternatively, the relativizer can be omitted in English: ‘*answers we’d like to have.*’ In these cases the TreeBank will still include a placeholder for the relativizer, but a 0 will appear where the explicit relativizer normally appears.

For example, in the following TreeBank annotation, the object position of the verb has a trace (NP \*T\*-6), which is being coindexed with the relativizer (WHNP-6 *that/which/0*).

*TreeBank annotation:*

```
(NP (NP answers)
  (SBAR (WHNP-6 that/which/0)})
  (S (NP-SBJ-3 we)
    (VP 'd
      (VP like
        (S (NP-SBJ *-3)
          (VP to
            (VP have
              (NP *T*-6))))))))))
```

Whereas, syntactically, the trace is being coindexed with the relativizer, semantically, there is a relationship between the trace and the NP *answers*, which is not being represented in TreeBank. This relationship is now captured via post-processing, so annotators do not need to provide any link here, but should be aware of how to recognize and understand the relationships between elements in a relative clause structure.

### 1.8.2 Reduced Relative Annotation

A relative clause may be reduced when passive, resulting in a unique syntactic structure. For example, a passive relative clause construction such as ‘*The woman that was dressed in blue walked past the house*’ can be reduced to ‘*The woman dressed in blue walked past the house.*’ Because the verb in these cases is always passive, the TreeBank annotation of reduced relatives will include an object trace after the verb. However, unlike normal passive constructions, this trace will never be coindexed with the subject. To show this relationship, an ‘&’ link is used between the object trace and the subject. Annotators should first annotate the object trace, noted by TreeBank as **none-\*** with the correct argument. Next, they should link this to the sister noun phrase containing the words represented by the object trace by pressing **Ctrl+Shift+7**.

*TreeBank Annotation:*

```
(S
  (NP-SBJ-1
    (DT This)
    (VP
```

```

(VBZ is)
(VP
  (VBN considered)
  (S
    (NP-SBJ
      (-NONE- *-1)
    (NP-PRD
      (NP
        (CD one)
      (PP
        (IN of)
        (NP
          (NP
            (DT the)
            (JJS biggest)
            (NNS caches)
          (VP
            (VBN seized)
            (NP
              (-NONE- *)
            (PP-LOC
              (IN in)
              (NP
                (DT the)
                (NN district)

```

*PropBank Annotation:*

REL: seized

ARG1: [NP -NONE- \*] & [NP the biggest caches]

ARGM-LOC: [PP-LOC in the district]

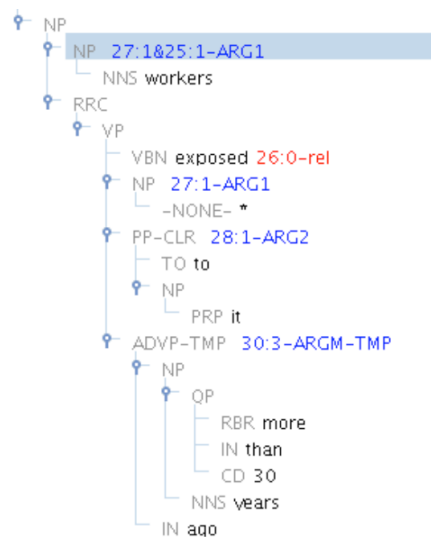


Figure 1.16: Example of a Reduced Relative annotation



## Annotation of PRO

Many traces found in the TreeBank arise as a result of the movement of a constituent from its canonical position. Movement leaves a trace, represented by a \* or a \*T\* in the TreeBank. PRO, on the other hand, does not arise as a result of movement. Rather, PRO arises where there is an underspecified, or unrealized subject of a verb. For example, the subject of the verb *leave* in the phrase ‘*she tried to leave*’ is not realized. However, the TreeBank will represent the unrealized subject of *leave* with PRO:

*She-1 tried \*PRO\*-1 to leave*

REL: leave

ARG0: \*PRO\*-1

In cases like that of the example above, the PRO element is already coindexed with the fully realized subject because the PRO is positioned in a clause that is governed by the higher clause with the same subject, *she*. Thus, the annotator need not add any additional links between PRO and the explicit subject. However, there are also cases in which PRO arises but it is not governed by a higher clause. In these cases, it is not coindexed with a fully realized subject. Figure 1.17 shows an example of a PRO which has no referent in this sentence.

When annotating PRO that is not indexed, if the annotator is certain that the subject is realized elsewhere in the instance, then a link should be created between PRO and the explicit reference. If the annotator is **not absolutely certain** that the explicit reference and PRO share the same referent, then the annotator should not create the link. Essentially, unless the relation is absolutely certain, we should err on the side of agreeing with the TreeBank annotation and its existing indices. In the cases where the annotator has decided with certainty that a link should be created between the PRO argument and a fully realized subject: first, annotate PRO with its appropriate argument, then select the node of the explicit subject mention associated with PRO, and finally, click **Argument** on the Jubilee menu bar, followed by clicking **Functions**. From the options therein, select \* (shortcut: **Ctrl+Shift+8**). At this point, the linked annotation should appear on the currently selected node of the explicit referent. An example of this is also given in Figure 1.18.

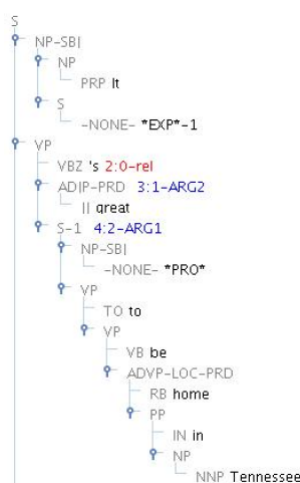


Figure 1.17: Example of a nonreferential *it* PRO

The goal of this annotation is to provide additional semantic information about the arguments of the verbs. In some cases, antecedents are not syntactic constituents, or have a different morphological form, as the possessive pronoun *your* below illustrates; forego linking in these cases:

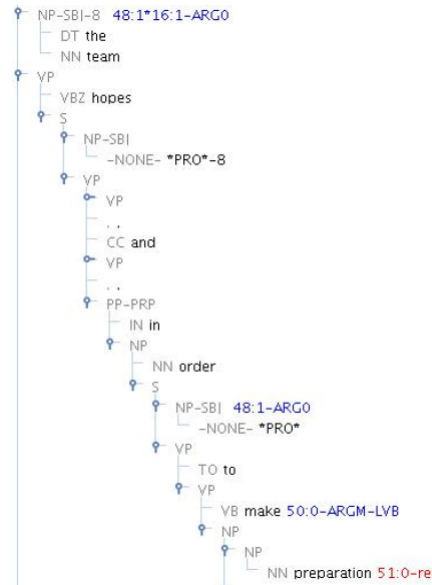


Figure 1.18: Example of PRO annotation

*On the issue of abortion , Marshall Coleman wants to take away your right [\*] to choose and give it to the politicians.*

ARG0: [\*PRO\*]

REL: choose

Additionally, note that the null element should be linked to the highest possible node containing its referent without recursively annotating other arguments or the **rel** itself (i.e. creating arguments embedded within other arguments). If this is not possible, the link should be omitted.

### 1.8.3 Concatenation of multiple nodes into one argument

As a rule, annotations should be placed on the highest node possible encompassing the entirety of a constituent (e.g., the NP or PP level in the TreeBank), and one tag should correspond to one node for every node within the appropriate annotation span. However, in some cases the **rel** is situated within an NP node, and the manner in which the TreeBank is laid out makes it impossible to capture a whole constituent under one node. In these cases, two leaves of the tree may have to be concatenated together under a single argument label. Figure 1.19 shows an example of this.



Figure 1.19: Example of a Concatenated Argument

To concatenate two leaves into a single argument, first select the node of the first word of the constituent (e.g. *1985*) and then click the appropriate button (e.g. 2) indicating which semantic role the argument is playing. Next, select the node of the second word of the constituent (e.g. *ruling*) and navigate to **Argument** on the Jubilee menu bar, followed by clicking **Functions**.

From the options therein, select `,` (shortcut: `Ctrl+Shift+,`). The resulting annotation will reflect the tree location of both nodes as part of a single argument (e.g. `3:0,4:0-ARG1`).

In cases of passive extraction of the subject and subject raising verbs like *seem*, concatenation can also be required to put the subject and clause following the verb under one argument label. See Section 1.9 for more discussion of this topic.

Another instance of concatenation involves verb particle constructions, such as *stalled out* in Figure 1.20. Many of these verbs will have a frame set especially for particle instances.

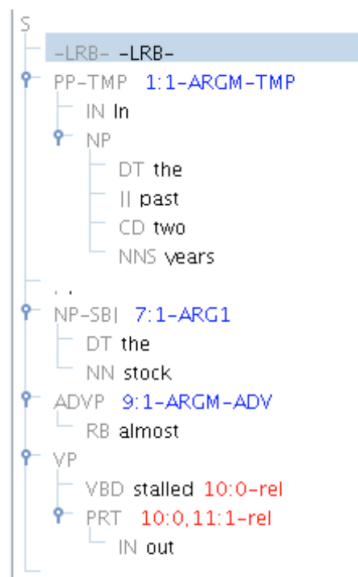


Figure 1.20: Example of a verb particle in Jubilee

#### 1.8.4 Special cases of topicalization

Topicalization occurs when a constituent is moved from its underlying syntactic position to an atypical position in order to draw attention to that constituent. Although the majority of cases of topicalization simply require annotation of a trace that is already indexed to the topicalized constituent, occasionally the topicalized constituent is repeated, often in the form of a pronoun, and the two constituents are not already indexed in the tree. For example, ‘*911, that’s the number you call in an emergency.*’ If the `rel` to be annotated is *is*, the annotator would firstly have to annotate the pronoun *that* as `ARG1`, then provide a coreference link to the pronoun’s referent, *911*. Performing this link works in an identical manner to that of creating coreference links for `PRO` seen in Section 1.8.2. Thus, after selecting and annotating the pronoun in the argumental position, subsequently select the topicalized node and click **Argument** on the Jubilee menu bar, followed by clicking **Functions**. From the options therein, select `*` (shortcut: `Ctrl+Shift+8`). The linked annotation should appear in the TreeBank view and in the annotation view at the top of the screen. For example:

*TreeBank Annotation:*

```
(S
  (NP-SBJ
    (PRP I))
  (VP
    (VBD expected)
```

```

(NP
  (PRP it))
(PP-LOC
  (IN in)
  (NP
    (NP
      (DT a)
      (NN country))
    (SBAR
      (WHNP-1
        (-NONE- 0))
      (S
        (NP-SBJ
          (PRP we))
        (VP
          (VBP love)
          (NP
            (-NONE- *T*-1))
          (, ,)
          (NP-TPC
            (PRP me)
            (CC and)
            (PRP you))))))))))

```

*PropBank annotation:*

REL: love

ARG1: (NP (-NONE- \*T\*-1))

ARG0: [NP-SBJ we] \*[NP-TPC me and you]

## 1.9 Special Cases: small clauses and sentential complements

This section is concerned with different types of clausal complements and modifiers, with a special focus on passive extraction, subject-raising verbs, and aspectual verbs. In the following sentence, the clause **S-CLR** has a trace in the subject position of *asleep*, which is coindexed with the subject of the verb *fell*, *I*.

*I fell asleep on the floor.*

*TreeBank annotation:*

```

S (NP-SBJ-1 I)
  (VP fell
    (S-CLR (NP-SBJ *-1)
      (ADJP-PRD asleep))
    (PP-LOC on
      (NP the lobby floor))))

```

When annotating the verb *fell*, the small clause (marked as **S-CLR** above) is tagged as **ARGM-PRD**, and the **ARG1** argument is the **NP-SBJ** *I*. Note that although the empty category **NP-SBJ \*-1** is being coindexed with *I*, the trace is not the argument of *fell*, but rather is the subject of *asleep*.

*PropBank annotation:*

REL: fell  
 ARG1: I  
 ARGM-PRD: [NP-SBJ \*-1] asleep

Verbs like *expect* are analyzed as having a clause as its argument (which corresponds to the event expected). In this case PropBank annotation follows TreeBank analysis of these sentences, where the clausal complement is being selected as ARG1:

*John expected Mary to come.*

*PropBank Annotation*

REL: expected  
 ARG0: John  
 ARG1: Mary to come

If such sentences are passivised, as shown below, then the ARG1 argument is the clausal complement of the verb. Parallel to ICH and RNR traces, we assume that the trace  $[-1]$  is being ‘reconstructed,’ so that the ARG1 in this case corresponds to the proposition ‘*Mary to come*’. It is necessary to annotate the dislocated portion (e.g. NP *Mary-1*) as part of the ARG1 via concatenation using the , operator, discussed in Section 1.8.3. The process by which the subject is raised from the clausal complement to become the subject of the matrix verb, *expect*, is sometimes called passive extraction.

*Mary-1 is expected [\*]-1 to come*

REL: expected  
 ARG1: [Mary-1] , [\*-1 to come]

A similar analysis applies to verbs like *seem* and *believe*, which are known as raising verbs. In PropBank annotation, the S clause and the dislocated argument are annotated as the ARG1 argument, again via concatenation, as in Figure 1.21.



Figure 1.21: Two Examples of Dislocated Arguments

And, finally, another class of verbs which follows this analysis includes aspectual verbs like *continue* and *start*, which take events as their arguments. Watch the rolesets for these verbs carefully, often there is a separation of the aspectual sense and the agentive sense of the verb (e.g., *begin.01* and *begin.02*).

[*New loans*]-4 *continue* [\*-4] *to slow*.

*PropBank* annotation:

REL: *continue*

ARG1: [[*New loans*]-4] , [\*-4 *to slow*]

## 1.10 Handling common features of spoken data

Annotation of transcripts of spoken data tends to be more difficult than annotation of spoken material due to disfluencies, repetitions and asides that do not normally occur in written English. Annotation procedures for each of these types of challenges are addressed in the following sections.

### 1.10.1 Disfluencies and Edited Nodes

Speakers often begin to say one utterance, stop due to a variety of speech errors or pragmatic factors, and then resume the utterance. Sometimes the speaker resumes with a very similar utterance, other times the speaker resumes with what seems to be an entirely different utterance. The Treebank handles such disfluencies with the use of a separate node, generally labeled **Edited**, such that the error portion of the utterance is separated from the remainder of the utterance within this node. If the relation is not within the edited node, simply ignore edited nodes and do not annotate them, regardless of whether or not they are within the span of annotation. If the relation is within the edited node, annotate in accordance with the normal span of annotation, but do not annotate anything past the cut off point of the utterance (this will often have a function tag in the TreeBank UNF, indicating that the utterance is unfinished). In other words, treat everything before the speaker stops and switches the progression of the utterance as you normally would any relation; however, do not annotate anything beyond where the speaker stops the original utterance and then starts the repaired utterance. See Figure 1.22 for an example of proper annotation.

This example brings to light several common challenges in annotating disfluencies. Notice in this example that the TreeBankers have assessed the portion of the utterance that is later repaired, and this ‘error’ portion is placed in an edited node. The argument of *be* prior to the cut off is annotated as usual. However, the beginning of the relative clause construction is not annotated; this is omitted because there is no index on the relativizer *that* as there normally would be, which is always linked to a trace somewhere in the rest of the utterance. As a result, there is no anchor within the clause for the relative clause construction, and it is therefore incomplete and shouldn’t be annotated. Conversely, if the trace linked to the relativizer were present within the edited node, it would require normal annotation. Although the entire instance is not shown here, it is also notable that there is no way for the annotator to know precisely which sense to use because the utterance is incomplete. When possible, try to use context to make the best guess of what sense of the verb is appropriate. It is often helpful to consider the repaired utterance after the cut off because speakers sometimes continue with a very similar utterance. However, in cases such as this one, where there is little relevant context to this portion of the utterance and the repaired utterance seems very different from the cut off utterance, simply select the most frequent sense of the verb. The most frequent sense of the verb should be the .01 sense; however, most of the .01 senses in PropBank were established during annotation of the Wall Street Journal. Thus, when tackling a very different corpus that is a transcription of spoken

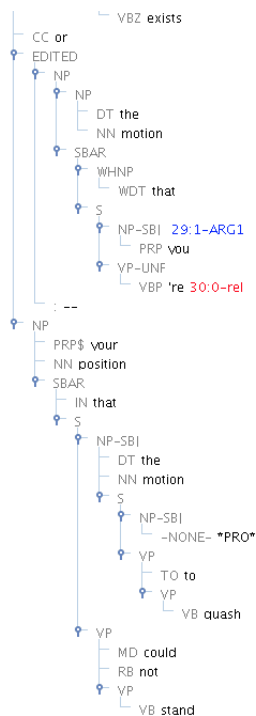


Figure 1.22: Correct annotation of disfluency in an edited node

data, it is possible that another sense of the verb will be particularly prevalent in that corpus. In these cases, annotators should use their best judgment of the patterns of that corpus to select what is the most likely sense of the verb in ambiguous cases.

### 1.10.2 Asides: PRN nodes

In both writing and perhaps somewhat more frequently in spoken discourse, speakers may insert an utterance that is not directly related to the main utterance in progress: *‘It is the livestock sector, according to a new report by the United Nations Food and Agriculture Organization, that generates more greenhouse gases than any other industry.’* In the preceding utterance, the *‘according to’* phrase would be encompassed in a node labeled PRN (meaning parenthetical) in the TreeBank. Like Edited nodes, PRN nodes should not be annotated unless the relation is within the parenthetical node itself. Annotation within the PRN node should be restricted to the normal span, meaning that it will not extend beyond the PRN node. In the case of spoken discourse, such nodes are often used for embedded phrases such as *‘I think,’ ‘You know,’* or repair initiators such as *‘I mean.’*

## Chapter 2

# Light Verb Annotation

Light verb usages are those which are considered semantically bleached, thus they do not carry the specificity of meaning that the verb would carry outside of a light verb construction (Butt, 2003). For example, a ‘heavy’ version of the verb *make* is used in the phrase ‘*She made a pie out of fresh cherries and refrigerated dough.*’ This usage reflects the normal semantic roles associated with the creation sense of *make*: Agent, Product and Material. A light version of the verb *make* is used in the phrase ‘*She made an offer to buy the company for 2 million dollars.*’ Unlike the first phrase, the verb *make* does not specify the semantics of the event; rather, the eventive noun or true predicate *offer* specifies the event. For this reason, we can often rephrase light verb constructions with the verb counterpart of the eventive noun (if one exists, this is not always the case) without losing the meaning of the utterance: ‘*She offered to buy the company.*’ In addition to specifying the semantics of the event, the eventive noun also projects the argument structure of the utterance. For example, the infinitival complement ‘*to buy the company*’ and the price argument ‘*for 2 million dollars*’ are canonical types of arguments for *offer*, but not for the verb *make*: ‘*\*She made to buy the company.*’ Similarly, the canonical arguments of the heavy sense of *make* are not compatible with *offer*: ‘*?She made the offer out of intentions, ink and paper.*’ Because the verb in these cases is not the element that specifies the semantics of the event or projects the argument structure, we cannot treat light verbs in the same way that we treat ‘heavy’ verbs. Therefore, we have special annotation procedures for light verbs, outlined in the following sections.

### 2.1 Pass 1: Verb Pass

Common light verbs in English are *make*, *take*, *get*, *give*, *have*, and *do*, found in light verb constructions such as ‘*John made an inspection of the premises,*’ ‘*John took a walk to the store,*’ ‘*John had a drink of iced tea,*’ and ‘*John did an investigation of the crime.*’ The verb *give* is often cited as a light verb in English as well; however, for the purposes of PropBank, *give* was not originally treated as a light verb because all light verb usages of *give* maintain the canonical transfer semantics and ditransitive argument structure: ‘*She gave him thunderous applause.*’ Nonetheless, for symmetry in the events identified in sentences such as ‘*She gave the baby a bath*’ and ‘*She took a bath,*’ current annotation procedures recognize *give* as a light verb where its semantics are overshadowed by that of its eventive or stative noun complement. Unlike *give*, most other light verbs do not maintain their canonical semantics or argument structure when used in light verb constructions. It is theoretically possible for many other verbs to be light in certain usages, such as ‘*She produced an alternation.*’ Whenever a verb seems to describe the event semantics less than the accompanying eventive or stative noun, it is likely a light verb construction. In particular, annotators should consider near synonyms of the light verbs (e.g., *have-possess*, *make-create*, *give-emit*, *get-undergo*, *do-execute*) to be light when they also



accompany an eventive or stative noun carrying the bulk of the semantics.

When a light verb is encountered, the annotator should perform the following steps:

*Step 1: Roleset selection*

Annotators should select the .LV roleset, available for all verbs, when the verb is found in a light verb construction. In some cases, placeholder numbered rolesets are available for light verbs. Do not use this numbered roleset, always use the .LV roleset.

*Step 2: Annotation*

Annotators should annotate only the eventive noun or true predicate itself as ARG-M-PRR (PReducating Relation). Unlike normal tag placement, the ARG-M-PRR tag should be placed directly on the lexical level or leaf containing the action nominal complement. Figure 2.1 illustrates correct annotation of light verb constructions:

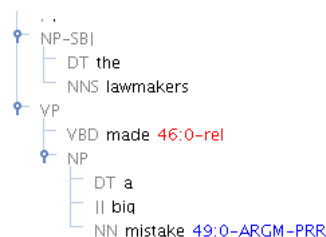


Figure 2.1: Correct annotation of a light verb construction

Annotate **only** this element. The first pass is strictly an identification pass meant to find and label light verbs as such. We do not annotate additional arguments because the verb itself is not the relation that is projecting the argument structure; thus it is not appropriate to annotate other arguments as if they were truly semantically related to the verb. Rather, the surrounding arguments are annotated during the second pass, when the action nominal complement is annotated, because it is the action nominal complement that projects the argument structure.

It is often difficult to decide if a certain usage is a light verb usage or not. Light verbs are thought to exist on a continuum ranging from the purely compositional meanings stemming from ordinary collocation of words (e.g. ‘*I tripped on the rug*’), wherein every word maintains its full semantic value, to the entirely non-compositional meanings stemming from fixed idiomatic expressions (e.g. ‘*I tripped the light fantastic*,’ meaning ‘*to dance*’). As a result, there are often fuzzy boundaries between heavy usages of verbs, light usages of verbs, and idiomatic usages of verbs. For the purposes of PropBank, annotators should be generous in their definition of light verb constructions and annotate accordingly when in doubt as to whether something is a light verb or not. In turn, cases where the annotator seems to have been too generous in this definition will be corrected in adjudication. Here are several short heuristics for when to annotate an instance as .LV:

1. Does the noun object denote an event or state? If not, it is not an LV. If so, proceed to the next question.
2. Are the arguments present more representative of typical arguments of the verb relation or the noun relation? (e.g. ‘*I made an offer to buy the house for \$200,000*’ — such price arguments are more typical of offering events than making events). If the arguments are more typical of the verb relation, it is not an LV. If the arguments are more typical of the noun relation, proceed to the next question. If you are unsure, here is a secondary test: try to add other common arguments from the verb’s rolesets to see if they are semantically felicitous with the usage in question. For example, a Material argument is typical of creating and making events; thus, one could test the appropriateness of a Material argument in the

usage: ‘?I made an offer out of paper and ink.’ If the addition of one of the verb’s arguments does not seem appropriate, this indicates that the semantics are likely stemming from the noun and the usage is light. If the addition of the verb argument does seem plausible, this is evidence that the semantics are stemming from the verb and it should be treated as a heavy usage.

3. Similarly, does the potential light verb link the eventive or stative noun to a subject argument that is typical of the noun relation, or the verb relation? If the subject argument seems to carry one of the typical semantic roles of the noun relation (e.g. ‘[I]<sub>AGENT</sub> made a call to the hospital;’ ‘[I]<sub>RECIPIENT</sub> got a call from the hospital’), it should be considered a light usage; proceed to the next question for a final check. If the subject argument does **not** carry one of the semantic roles typical of the noun relation (even if the other arguments present do seem to be typical of the noun relation), then this is evidence that the subject’s semantic role has been assigned by the verb, and the usage should not be treated as light.
4. If you are not sure of the answers to (2) and (3), consider rephrasing the verb + noun combination using a lexical verb related to the noun if one exists (to continue the previous example, ‘I called the hospital;’ ‘I was called by the hospital’). If the rephrasing still captures the majority of the event semantics (even if there are small variations in voice, aspect, or tone), then mark the instance as a light verb, as the rephrasing provides evidence that the surrounding elements in the sentence are projected by the semantics of the noun even (which you’ve just rephrased as a verb) rather than the verb. If the rephrasing results in loss of meaning, then the verb must add semantics beyond what is captured by the eventive or stative noun should it be considered heavy. For example, ‘I attempted to call the hospital’ loses some meaning with the rephrasing ‘I called the hospital.’ This test is helpful, but a semantically similar lexical verb may not exist, so one must ultimately rely on considering whether the arguments ultimately stem from either the noun or the verb.

These heuristics are summarized in the following flow chart.

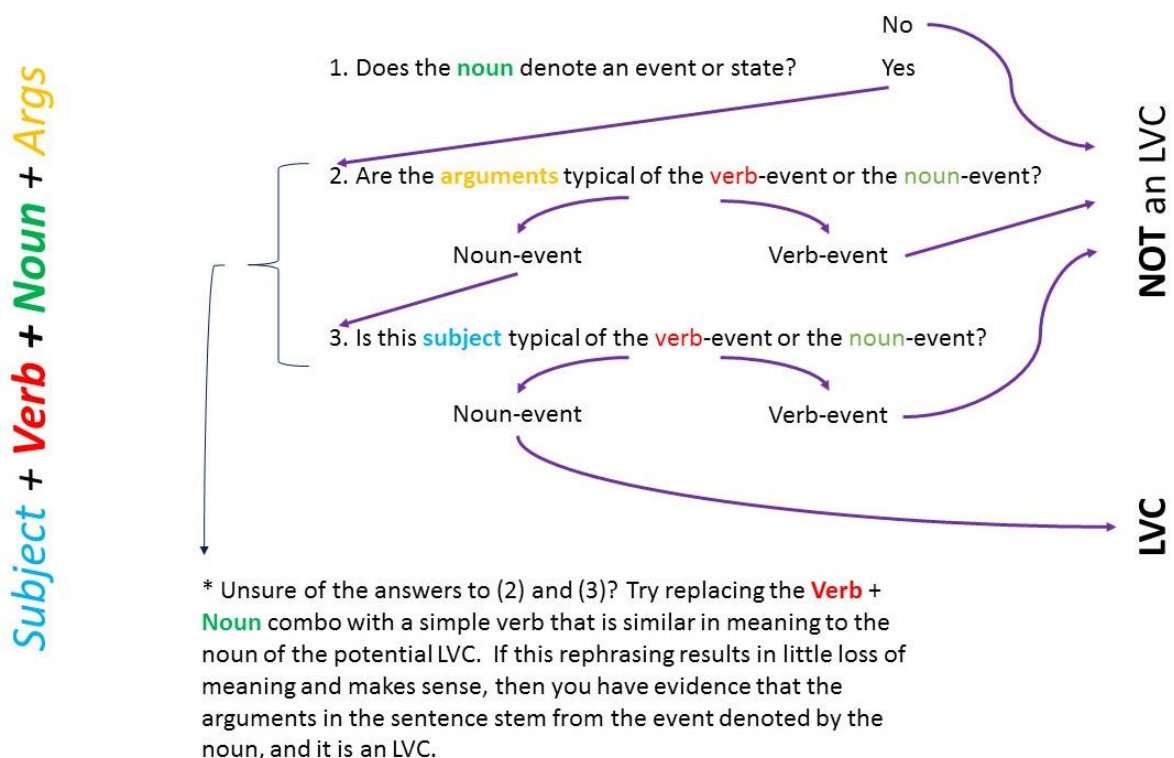


Figure 2.2: Light Verb Heuristics Flow Chart

## 2.2 Pass 2: Noun Pass

The bulk of the annotation of light verb constructions will be performed during the second noun pass, wherein the action nominal complement or true predicate is the relation. Unlike ordinary noun annotation described in Chapter 3, annotation of nouns that are true predicates within a light verb construction requires the annotation of arguments within both the noun relation's span of annotation (sisters to the noun and sisters to the noun phrase), and the light verb's span of annotation (sisters to the verb and sisters to the verb phrase). In the case of light verb constructions, the syntactic arguments of both the light verb and the action nominal complement are annotated together because it is thought that both contribute to complex predication. After identifying the fact that the noun relation is a true predicate in a light verb construction, proceed with the following steps:

### *Step 1: Roleset Selection*

Roleset selection proceeds exactly as it does for normal annotation of noun relations: select the appropriate numbered roleset according to the sense of the usage.

### *Step 2: Annotation*

Annotate direct arguments of the noun relation (sisters to the noun and sisters to the noun phrase) and the syntactic arguments of the light verb (sisters to the verb and sisters to the verb phrase) in accordance with the argument structure outlined in the selected roleset. In addition, annotate the light verb itself using the ARGM-LVB tag, place this tag directly on the leaf node of the light verb. For example:

*Yesterday, Mary made an accusation of duplicity against John because she was enraged with jealousy.*

ARGM-TMP: Yesterday

ARG0: Mary

ARGM-LVB: made

REL: accusation

ARG2: of duplicity

ARG1: against John

ARGM-CAU: because she was enraged with jealousy.

Thus, all arguments of the complex predicate (in this case make+accusation) are annotated in accordance with the true predicate's argument structure, outlined in the noun relation's roleset.

## 2.3 Examples

Here is a list of examples of less typical light verb constructions:

1. *He got a sense of his soul.*

ARG0: He

ARGM-LVB: got

REL: sense

ARG1: of his soul

2. *He does have a prior conviction for the knifepoint rape.*

ARG1: He

ARGM-LVB: have

ARGM-TMP: prior

REL: conviction

ARG2: for the knifepoint rape

3. *Officials got a tip ICH-1 back then [that someone may have dumped the equipment]-1.*  
 ARG1: Officials  
 ARGM-LVB: got  
 REL: tip  
 ARG2: \*ICH\*-1 (that someone may have dumped the equipment)  
 ARGM-TMP: back then
4. *\*PRO\* have some fun.*  
 ARG0: \*PRO\*  
 ARGM-LVB: have  
 REL: fun
5. *You get the US attention.*  
 ARG1: You  
 ARGM-LVB: get  
 ARG0: US  
 REL: attention
6. *I had no right to print that.*  
 ARG1: I  
 ARGM-LVB: had  
 ARGM-NEG: no  
 REL: right  
 ARG2: to print that
7. *He took pride in his hiding.*  
 ARG0: He  
 ARGM-LVB: took  
 REL: pride  
 ARG1: in his hiding
8. *We get a different excuse for this every time.*  
 ARG3: We  
 ARGM-LVB: get  
 ARGM-ADJ: different  
 REL: excuse  
 ARG1: for this  
 ARGM-TMP: every time
9. *Charter schools are having some successes in this area.*  
 ARG0: Charter schools  
 ARGM-LVB: having  
 REL: successes  
 ARG1: in this area
10. *China gives that impression to the world.*  
 ARG0: China  
 ARGM-LVB: gives  
 REL: impression  
 ARG1: to the world

Each of the instances given above could be considered a somewhat ‘unusual’ LVC, and therefore somewhat difficult to identify, for the following reasons: 1) the eventive or stative noun may not have a clear verbal counterpart, so it is hard to rephrase with a lexical verb to get a sense of what the event really is, 2) the subject argument linked by the verb is not an agent, and may

be something as peripheral as a listener or recipient, 3) the light verb may seem to change the event’s valency, similar to a passive voice construction. Despite these unique characteristics, each of the above instances should be annotated as an LVC.

## 2.4 Tricky Cases

There are also tricky cases of verbs being used in an abstract way with a noun complement that is eventive or stative that should **not** be annotated as LVCs. Generally, these can be distinguished from LVC cases because the subject argument linked by the verb does **not** denote one of the eventive or stative noun’s normal semantic roles. Here are a few examples:

1. *We’ve **had assassinations** before this.*

Note: Here, *We* does not denote the person assassinated or the killer, but rather just the group of people affected by the assassinations. The best roleset for this case would be **have.03**, indicating abstract possession.

2. *\*PRO\* **give** us your **assessment**.*

Note: Here, the semantics of transfer overshadow the semantics of *assessment*, as indicated by the fact that the recipient, *us*, does not denote an argument that’s part of the normal *assessment* roleset. Because the transfer semantics are strong for the above example, the best roleset would be **give.01**. For light cases of *give*, what can be thought of as the Recipient in terms of verb semantics is often the patient in terms of the eventive or stative noun semantics: ‘*He gave his opponent a quick jab,*’ i.e., ‘*He quickly jabbed his opponent.*’ In other cases, the Recipient in terms of verb semantics is also a recipient in terms of the eventive or stative noun’s semantics, but importantly, in these cases the noun’s roleset includes a recipient of some sort, such as a listener: ‘*The president gave a rousing speech to the class of 2010.*’

3. *It **gives** you a **sense** of the terror she must have felt.*

Note: Here, *give* introduces an outside causer that is not part of the eventive noun’s roleset (something is causing you to sense something). Since a sense of what she felt is being metaphorically transferred or imparted, the best roleset is **give.01**. Similarly, ‘*You gave me the ride of my life*’ introduces an outside causer of the *ride* that is not part of the normal *ride* roleset.

The common thread here is that although the verb is potentially light and is accompanied by an eventive or stative noun, it should not be annotated as an LVC because the verb is not acting as part of a complex relation to syntactically link the eventive or stative noun to the subject argument.

The usages of *give* that introduce an outside causer bring up the related fact that verbs whose lexical semantics include causation, such as *cause* or *make* should also not be annotated as light verbs, even where they happen to link an eventive or stative noun complement to a potential Agent/Cause ARG0: ‘*The failing economy caused the end of the industry.*’ Although in many ways this case does fit the definition of a light verb (‘*the failing economy*’ can be seen as an agent of *end*: ‘*The failing economy ended the industry*’), because causation is part of the primary semantics of the verb *cause*, it is more consistent to consider these arguments to be primarily projected by the verb rather than the eventive noun, and therefore forego annotation as a light verb.

## Chapter 3

# Noun Annotation Instructions

Nouns can also be relations in the same way that verbs are often relations. Nominalizations and eventive nouns are especially likely to have an argument structure similar to that of a verb. For example, ‘*The troops’ destruction of the city,*’ can be thought of as the nominal counterpart of the clause ‘*The troops destroyed the city.*’ Thus, PropBank treats certain nominalizations and eventive nouns as relations to be annotated in the same manner as verb annotations: 1) select the appropriate roleset created specifically for the noun relation, 2) annotate numbered arguments in accordance with this roleset, and 3) annotate modifier arguments with the appropriate ARGM tags. However, there are a few differences to be aware of when annotating nouns, which are outlined in the following sections.

### 3.1 Span of Annotation

The span of annotation for noun relations mirrors that of verb relations. Instead of annotating the sisters to the verb and sisters to the verb phrase, annotation is required for the sisters to the noun and sisters to the noun phrase that encompasses the noun relation. Figure 3.1 illustrates the noun relation’s span of annotation. Also like the verb span of annotation wherein several VP nodes may encompass a verb, several NP nodes may encompass the noun relation, and arguments of the noun relation to be tagged can be sisters to each of these NP nodes. In the case of verbs, the annotator can use the position of the clause boundary to determine where the span of annotation is delimited. However, in the case of nouns, the boundary of the span is marked by another NP (or occasionally NOM) node. Annotators must determine which NP node is the stopping point by finding which is a sister to the verb. Annotators cannot annotate sisters to the highest NP node that is a sister to the verb because this will entail annotating sisters to the verb itself. This is illustrated in Figure 3.2. The long arrows indicate valid arguments of the noun relation, while the top T indicates the sister to the highest NP node, which is a sister to the verb *is*, thus, annotators should consider this NP node as the stopping point for annotation.

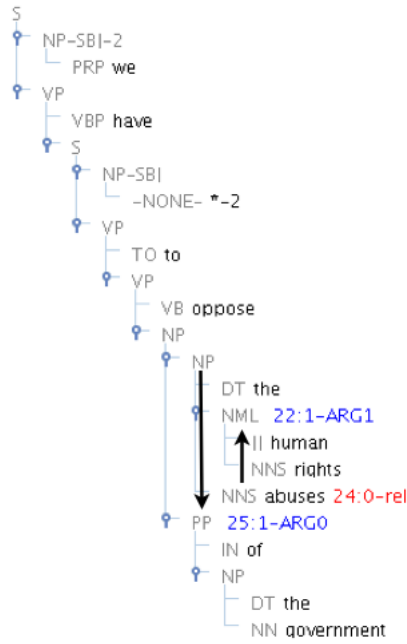


Figure 3.1: Noun span of annotation: the arrow to the right points to the sister of the noun, and the arrow to the left points to the sister of the noun phrase encompassing the noun relation

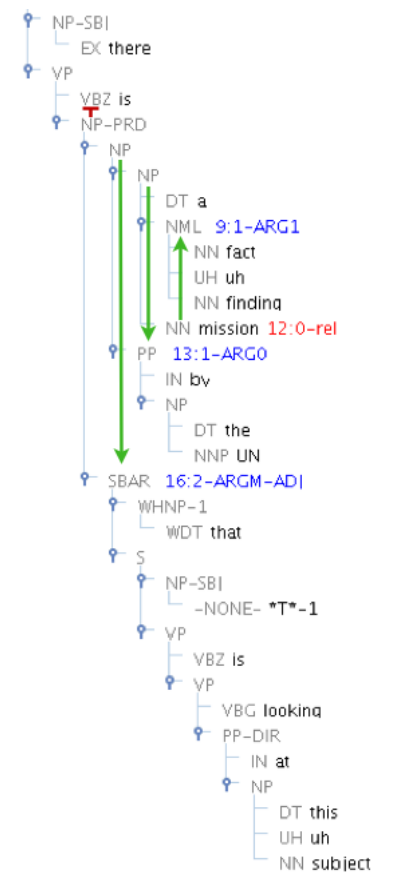


Figure 3.2: Noun span of annotation: the arrows point to arguments of the noun relation that should be tagged, while the highest NP node with the ‘T’ above it indicates the delimiting node of the span

In the case of a list of nouns ([NP and NP], [NP but not NP], [NP, NP, and NP], and so on), the span of annotation does not extend to list items other than the one containing the Rel. Figure 3.3 shows an example of correct list annotation.

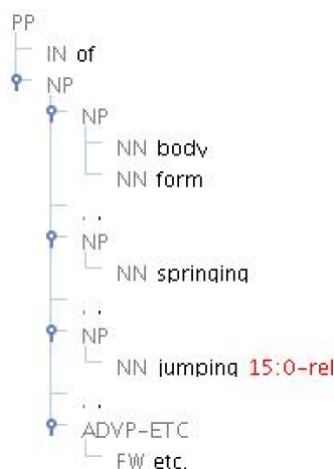


Figure 3.3: Noun List Annotation

### 3.2 Annotation of Numbered Arguments

Numbered arguments for noun relations are outlined in the corresponding noun frame file and rolesets therein. These should be selected in Jubilee in the same manner as verb rolesets. It is important to keep in mind that the type and order of constituents is highly variable for noun relations. For example, both agents and patients can appear before the relation as a possessor or noun modifier, or in a prepositional argument after the relation.

*Channel Nine's broadcast of the nightly news was praised for its quality.*

ARG0: Channel Nine's

REL: broadcast

ARG1: of the nightly news

*The nightly news broadcast of Channel Nine was praised for its quality.*

ARG1: nightly news

REL: broadcast

ARG0: of Channel Nine

As a result of this variability, it is especially important to think of the primary goal of PropBank when annotating noun relations: assign the appropriate semantic tag across various syntactic realizations of the argument.

### 3.3 Annotation of Modifiers

Modifier arguments are identical for nouns with one exception, the ARG-M-ADJ tag described below. That which is often expressed as an adverb in the case of verb relations is expressed with an adjective for noun relations:

*Was your personal experience of having been on the cruise pleasurable?*

ARG0: your

ARG-M-MNR: personal

REL: experience

ARG1: of having been on the cruise



In the above example, one can think of *personal* as equivalent to *personally* in the clause counterpart of the phrase: ‘*You personally experienced having been on the cruise.*’ If it is helpful, try to rephrase the instance in this manner to decide what type of argument it would be when accompanying a verb, and it will be the same type of modifier when accompanying the noun relation.

### 3.3.1 Adjectival modifiers (ADJ)

Instead of the modifier tag ADV, which should never be used for noun relations, the modifier tag ARGM-ADJ is used to label arguments that cannot be appropriately labeled with any other ARGM tag. As with the ARGM-ADV tag, the ARGM-ADJ tag should only be used as a last resort, when no other argument label can possibly fit. For example:

*The mayor’s shocking abuse of public funds outraged citizens.*

ARG0: The mayor’s

ARGM-ADJ: shocking

REL: abuse

ARG1: of public funds

In the above example, the manner of the abuse is not shocking, ‘*he shockingly abused public funds;*’ rather, the entire event is perceived as shocking by outsiders. Its verb counterpart would be something akin to ‘*Shockingly, the mayor abused public funds,*’ and would be annotated as ARGM-ADV for lack of a better tag. Similarly, modifiers such as this must be tagged as ARGM-ADJ for lack of a more specific tag. SBAR modifiers, as seen in Figure 3.2 should also be tagged as ARGM-ADJ, along with quantity phrases, such as Figure 3.4.



Figure 3.4: An Example of Adjectival Quantity

### 3.3.2 Secondary Predication modifiers (PRD)

In noun annotation, verbal modifiers of the **rel** are considered secondary predication because they are eventually annotated as well. Figure 3.5 shows two instances of secondary predication in noun annotation. An important distinction in ‘*...a judgement based on the law*’ is that the VP is only included in the span of annotation because of the NP mother node encompassing both the NP containing the **rel** and the VP itself. If the mother NP was an S node, the VP would not be included in the span of annotation.

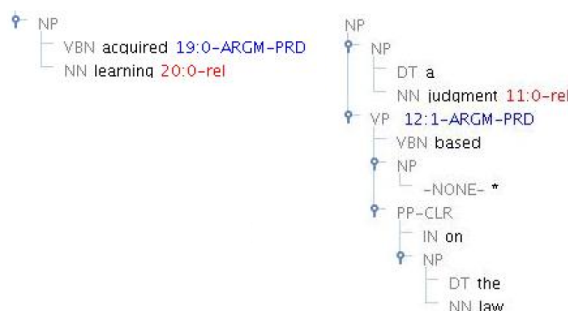


Figure 3.5: An Example of Secondary Predication in Noun Annotation

### 3.4 Exceptions to Annotation: Determiners and Other Noun Relations

As with verb annotation, the articles *a*, *an*, and *the* should not be annotated when they are in the span of annotation for a noun relation, nor should the majority of other determiners (*this*, *that*, *some*, *any*, etc.), labeled DT in the Treebank. The greatest exception to this is, of course, possessive determiners which frequently correspond to numbered arguments. Additionally, certain determiners such as *first* and *final* can be annotated as ARGM-TMP when they denote the time or frequency of the event. Finally, negative determiners *no*, and *neither* should be tagged as ARGM-NEG.

Just as we do not annotate other coordinated verbs that fall within the span of annotation of a given verb relation, we do not annotate other coordinated eventive nouns or nominalizations that fall within the span of annotation of a given noun relation. For example:

*Mary's investigation and eventual condemnation of the local government made the news.*

ARG0: Mary's

ARGM-TMP: eventual

REL: condemnation

ARG1: of the mayor

Outside of these exceptions, all other arguments within the span of annotation should be annotated.

### 3.5 .YY Roleset

In some instances, there are nouns which come up for annotation which should not be annotated. These cases include:

#### 1) Non-eventive nouns

A noun must denote an event or a state. Eventive nouns are always abstract, not concrete, but not all abstract nouns are eventive. For example, *building* in the sentence ‘*The building on the left...*’ refers to a concrete entity and should not be annotated, while *building* in the sentence ‘*The building of the Eiffel Tower many years ago*’ refers to an abstract event and should be annotated. Figure 3.6 shows further examples of the difference.

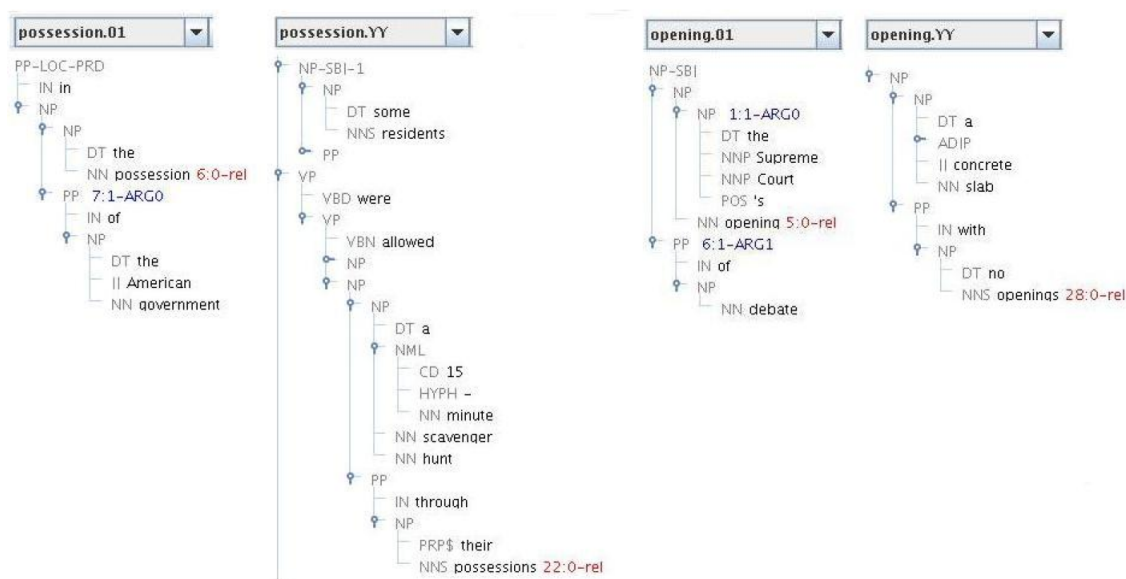


Figure 3.6: An Example of Non-Eventive Nouns (.YY) and Eventive Nouns

## 2) Non-head nouns

A noun **rel** must be a head, or a final noun in any noun-noun compound. The head noun is the noun which projects argument structure. Any noun which comes up for annotation which is not a head noun should not be annotated, but marked with the **.YY** roleset. Figure 3.7 shows an example of a head versus non-head noun **rel** distinction. In the instance marked as **.YY**, the **rel** *jumping* is not the head noun; the noun *jumping* modifies the head of the noun phrase, *portion*. The instance marked **learning.01** shows a head noun in an unusual position (where a modifier is expected).



Figure 3.7: An Example of Head Nouns and Non-Head Nouns (.YY)

## Chapter 4

# Adjectival Predicate Annotation Instructions

Crosslinguistically, it is common for there to be overlap between what is expressed as a verb and what is expressed as an adjective. Even in English, it can be difficult to distinguish between copular constructions and auxiliaries (e.g. ‘*He is limited/balding*’). In other languages, what would be considered an adjective in English may be expressed as a verb (given evidence from verbal morphology); for example, in Lakota, ‘*I am thirsty*’ would be expressed with the stative verb *imapuze*. Because PropBank is in part a resource for machine translation and several parallel PropBanks exist in different languages, it is important to annotate predicate adjectives in English.

### 4.1 Span of Annotation

The span of annotation for adjectives is very similar to that of the noun pass in LVC annotation. Because nouns require the support of a verb to express arguments, the predicate adjective and support verb together form what can be thought of as a complex predicate. Unlike LVC annotation, there is no special tag for the copular, or support, verb itself. However, annotation is required for sisters of the adjectival predicate (marked JJ in the Treebank) itself, sisters to the adjectival phrase that contains the predicate (ADJP), sisters to the support verb (with the exception of the ADJP sister itself; tagging this will cause recursive annotation since the ADJP contains the JJ and potentially other arguments), and sisters to the VP(s) that contains the support verb. An image of correct adjective annotation is given in Figure 4.1.

### 4.2 Annotation of Arguments

Annotation of numbered arguments follows the adjectival predicate’s roleset, as in both noun and verb annotation. Annotation of ARGMs proceeds in exactly the same manner as verb annotation (not noun annotation: the ADJ and LVB tags will not be used; the ADV tag is used). Additionally, exceptions to annotation (e.g., determiners or coordinated adjectival predicates) will also be followed as outlined for verb annotation.

### 4.3 Annotating Constructions

Essentially all gradable adjectives can participate in degree and comparative patterns that license the adjective relation to take (an) extra argument(s). As evidenced by the fact that these patterns are generalizable across adjectives rather than internally dictated, arguments are likely

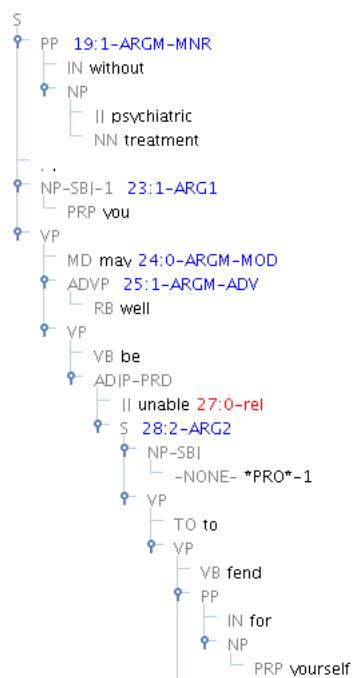


Figure 4.1: Adjective Span of annotation: note that sisters to both the JJ predicate and the support verb are both annotated.

projected by the constructions themselves, rather than the relation. In order to show this pattern, the arguments of the construction should be annotated with the Construction (CXN) marker. These CXN markers should be concatenated if and when the construction requires both parts or arguments to complete the pattern. Currently, three frequent constructions have been identified occurring with adjective relations: the Comparative Construction, the Degree Construction, and ‘the Xer the Yer’ Construction. These constructions are discussed in turn below. It is, of course, possible that other annotatable constructions exist. If it is noted that there is an argument present that is not normally projected by the semantics of the adjective relation, and instead there is a patterned phrasing (or construction) surrounding the relation that is generalizable to a variety of other relations, then annotators can use the CXN tags to mark the constructional argument. In these cases, please take a screen shot of the potential construction and contact the annotation supervisor with details.

### 4.3.1 Comparative Constructions

Comparative sentences fit a ‘more/less/as-X than/as Y’ pattern. For example, in ‘*Hillary Clinton is **about as** damaging to the Dem Party **as Jeremiah Wright**.*’ Hilary Clinton is compared to Jeremiah Wright in terms of damage done. Correct annotation of this instance is shown in Figure 4.2.

All participants in the construction are tagged with the CXN marker here. Note that while both *about* and *as* are required to tell the degree to which she is damaging, they are not subsumed under one node. Regardless of the number of CXN markers needed to capture the entire pattern, all tags should be concatenated (see Section 1.8.3 to review how to concatenate arguments if necessary). In some cases, the first comparison word (*more/less*) is lexicalized as part of the adjective relation: *better*, *worse*, *taller*, *smarter*, etc. In these cases, only the second portion of the construction (the *than* phrase) is marked with the CXN tag. Finally, if only the first portion of the construction is given, without a full mention of what something is being compared to, no CXN tag will be used, as there is really no argument projected by the construction in this case

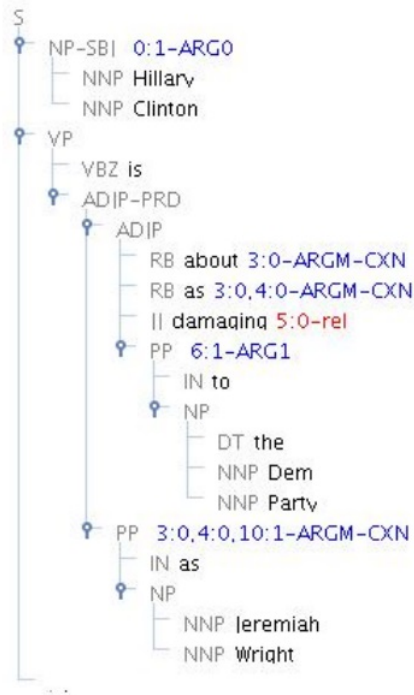


Figure 4.2: Annotation of a Comparative Construction.

(e.g. ‘Someday, things will be better’).

#### 4.3.2 Degree Construction

Constructions of this type tell the degree to which something is done or a state is true, and often mention a consequence of the degree to which a state is true. These sentences complete the pattern ‘X is too/so Y to/that Z,’ or ‘X is not Y enough to/that Z.’ Annotation is done the same way for degree sentences as for comparatives, with both the degree word and its counterpart (generally an infinitival or complement clause) receiving the concatenated CXN marker. For example, in Figure 4.3, both the degree indicator *too* and its descriptive counterpart ‘*\*PRO\* to learn*’ are concatenated with the CXN marker:

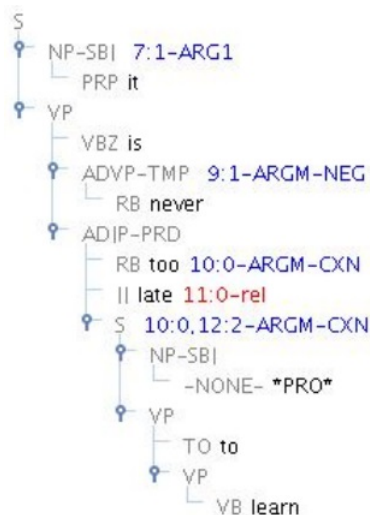


Figure 4.3: Annotation of a Degree Construction.

### 4.3.3 The Xer the Yer Construction

A variation of the comparative construction, these constructions fit the pattern ‘the X-er the Y-er.’ When X and Y are concatenated here, we see a resulting state; for example, in ‘*the longer he is around the more miserable I will be*,’ we are able to show that his staying longer results in a more miserable state. Correct annotation of this construction is shown in Figure 4.4.

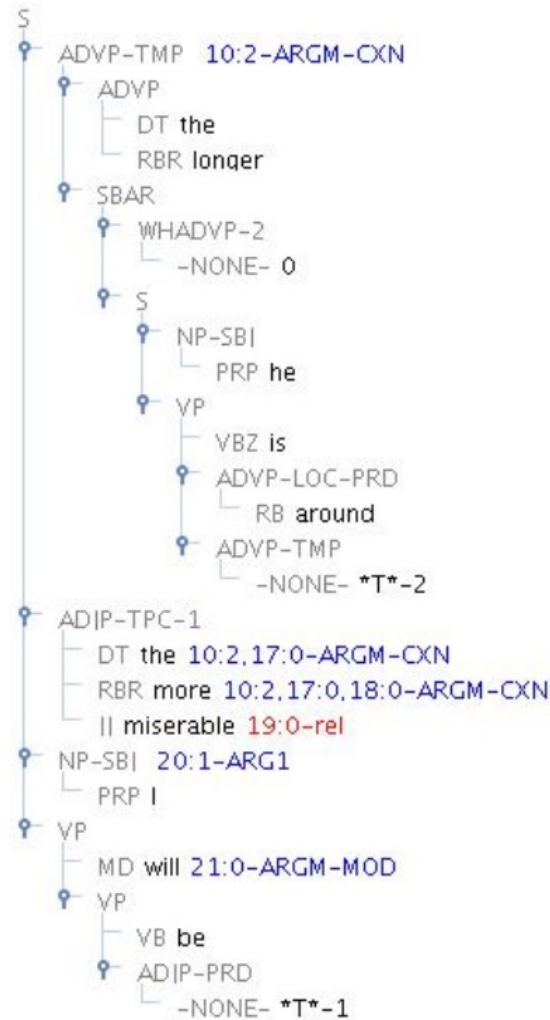


Figure 4.4: Annotation of a ‘The Xer the Yer’ Construction.

These patterns will almost always require at least three CXN markers. Also, note that the determiner *the* is considered to be an integral part of this construction.

# Appendix A

## Jubilee Hotkeys

These tables illustrate Jubilee’s Hotkeys, along with Section references for each function.

Basic Argument Annotation		
Function	Hotkey	Reference Section
rel	Ctrl+Shift+ /	NA
Argument 0	0	Section 1.3.1
Argument 1	1	Section 1.3.1
Argument 2	2	Section 1.1
Argument 3	3	Section 1.1
Argument 4	4	Section 1.1
Argument 5	5	Section 1.1
Argument A	A	Section 1.3.2
Comitative	8	Section 1.4.1
Locative	L	Section 1.4.2
Directional	D	Section 1.4.3
Goal	G	Section 1.4.4
Manner	M	Section 1.4.5
Temporal	T	Section 1.4.6
Extent	E	Section 1.4.7
Reciprocals	9	Section 1.4.8
Secondary Predication	7	Section 1.4.9
Purpose	P	Section 1.4.10
Cause	C	Section 1.4.11
Discourse	I	Section 1.4.12
Adverbials	V	Section 1.4.16
Adjectival	J	Section 3.3.1
Modal	O	Section 1.4.13
Negation	N	Section 1.4.14
Direct Speech	S	Section 1.4.15
Light Verb	B	Section 1.4.18
Erase Annotation	-	NA

Table A.1: Common Jubilee Hotkeys



<b>Concatenation, Null Elements, Traces</b>		
<b>Function</b>	<b>Hotkey</b>	<b>Reference Section</b>
Link Null Element with Coreferential Argument	Ctrl+Shift+8	Sections 1.7, 1.8
Concatenate Arguments	Ctrl+Shift+,	Sections 1.7, 1.8
Link Unindexed Object (Reduced Relative only)	Ctrl+Shift+7	Sections 1.7, 1.8
Reset Rel or Concatenate Verb Particles	Ctrl+Shift+/ 	Sections 1.7, 1.8

<b>Toolbar and Other Functions</b>		
<b>Function</b>	<b>Hotkey</b>	<b>Reference Section</b>
File > Open	Ctrl+O	NA
File > Save	Ctrl+S	NA
File > Quit	Ctrl+Q	NA
Treebank > Previous Tree	Comma	NA
Treebank > Next Tree	Period	NA
Treebank > Jump to ...	Ctrl+J	NA
Treebank > View Tree in Text	Ctrl+T	NA
Treebank > View Tree in Context	Ctrl+Shift+O	NA
Frameset > Previous Roleset	[	NA
Frameset > Next Roleset	]	NA
Frameset > View Examples	Ctrl+E	NA
Frameset > View Arguments	Ctrl+W	NA
Frameset > View Roleset Comments	Ctrl+C	NA
Help > About	F1	NA

Table A.2: Concatenation, Null Elements, Traces, and Toolbar Function Hotkeys

## References

- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the Linguistic Annotation Workshop*.
- Miriam Butt. 2003. The light verb jungle. In C. Bower G. Aygen and C. Quinn, editors, *Papers from the GSAS/Dudley House Workshop on Light Verbs.*, pages 1–50. Cambridge, Harvard Working Papers in Linguistics.
- Jinho D. Choi, Claire Bonial, and Martha Palmer. 2009. Jubilee: Propbank instance editor guideline (version 2.1). Technical Report 01-09, Institute of Cognitive Science, the University of Colorado at Boulder.
- David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(1):547–619.
- Beth Levin and M. Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface; Linguistic Inquiry Monograph*. MIT Press.
- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.