

Baby SRL Phase Two: Specs 1.5

Lori Moon, Cindy Fisher, Sandra Franco

Draft of 10/26/2015

1. Project Background and Goals

This project is the second major phase of work annotating adult-child interactions. The goals, methods, and publications of the first phase can be found at

http://cogcomp.cs.illinois.edu/page/resource_view/24

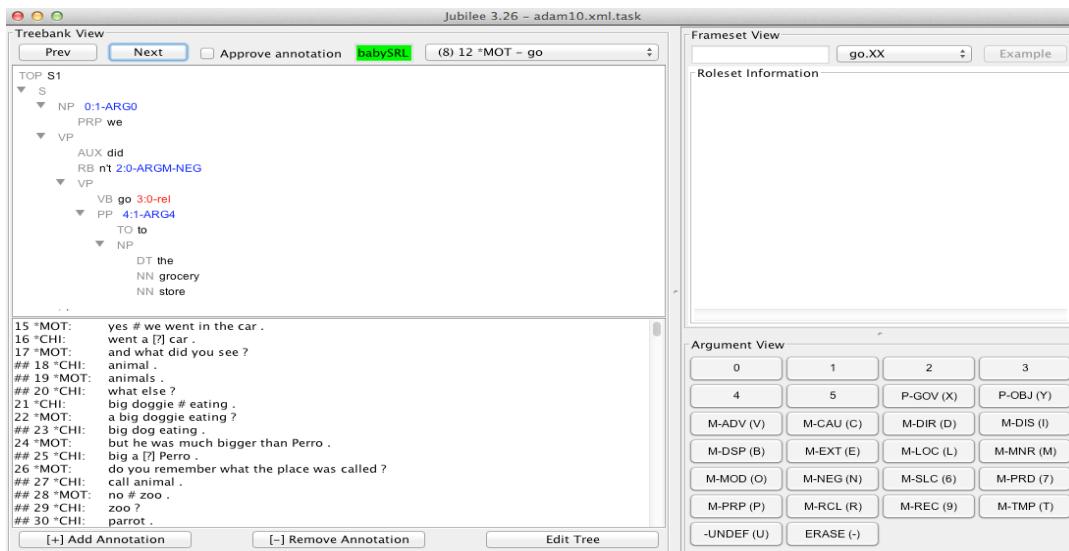
In this second phase, two major expansions are made to the scope of the data annotated. The first expansion is the addition of the annotation of child utterances and adult fragments that contain predicates. In the previous phase, only complete adult utterances were annotated.

The second expansion is the annotation of the semantic roles of the governor and object of prepositions (Srikumar & Roth 2013). Furthermore, in the first BabySRL annotation phase, main-verb uses of the copula, 'to be', were not annotated, so sentences such as "What's that" received no verb annotations. In the second phase, we annotate these uses based on Propbank guidelines. The scope is also more narrow, in that we only do these more in-depth annotations for the Adam files 1-23 -- not Eve or Sarah.

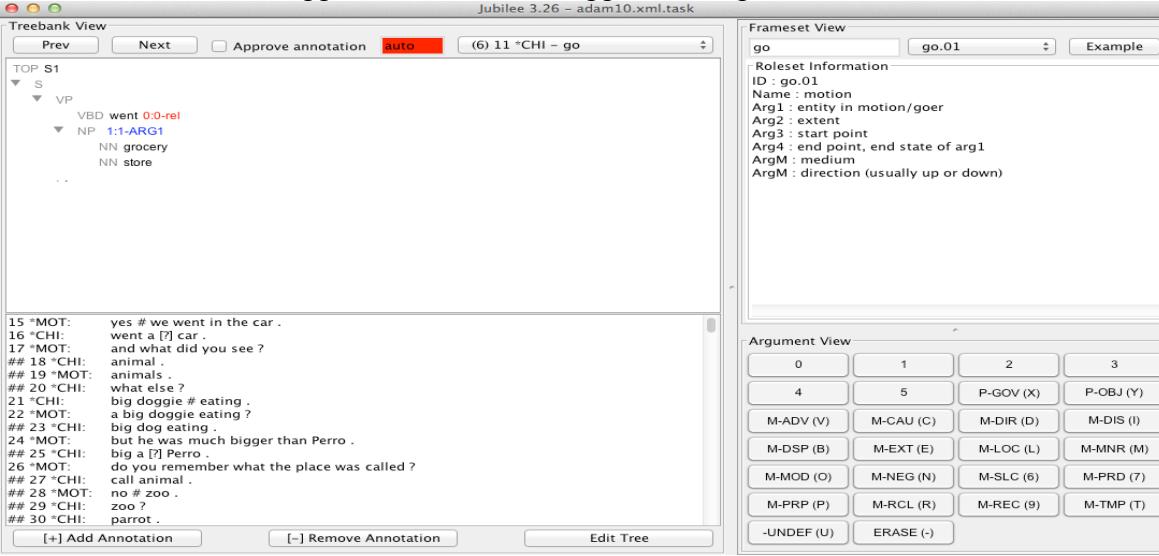
1.1. Annotation Methods

Annotators are presented with one parsed sentence at a time in the Jubilee annotation tool. The sentences appearing in the tool are either sentences annotated in the previous phase of the Baby-SRL project or sentences that were automatically parsed using the Charniak parser (Charniak & Johnson 2005).

Data samples that were annotated in phase one of the project are labeled "babySRL", and the label appears in green in the upper center portion of the window, as shown below:



The sentences which were parsed for phase two using the Cognitive Computation Group tools were labeled "auto", and the word appears in red in the upper center portion of the tool, as shown below:



Semantic role labels were automatically assigned to the parsed sentences using CogComp's illinois-srl (Punyakanok et al. 2008) and illinois-prep-srl tools (Srikumar & Roth 2013).

As a general guideline, parses are not changed any more than is necessary to have the semantic role labels at the right height in the tree and on the right type of constituent structure.

1.2. Jubilee Annotation Tool

1.2.1. Propbank Semantic Roles

Propbank semantic roles appear in the Jubilee tool window. The "Frameset View" provides a list of Propbank senses that the annotators can choose among, either if the existing annotation is wrong or if adding annotations.

1.2.2. Context Window

Each sentence appears in the tool with a context window. The context window is consulted by annotators when seeking to resolve ambiguities in sentences that could be cleared up easily, given the surrounding discourse.

1.2.3. Gold Mode

The Jubilee Gold mode allows two annotators to compare their annotations and adjudicate regarding the best parse and semantic role labeling.

1.2.4. Syntactic tree editing

Modifications to the Jubilee tool allow editing of syntactic trees for each annotation. A reference list of licit terminal and non-terminal symbols used in the PTB has been added to the tool as well.

An example of the gold mode is below:

The screenshot shows the Jubilee 3.26 software interface. The left panel, 'Treebank View', displays a tree diagram for a sentence from adam06.xml.task. The root node is 'TOP S1'. Below it is an 'INTJ' node containing the word 'UH yeah'. This is followed by an 'S' node, which contains an 'AUX do' node (with 'NP 2:1-ARG0 PRP you'), a 'VP' node (with 'VB find 3:0-rel' and 'NP 4:1-ARG1 PRP it'), and several other nodes. A large block of annotations is shown at the bottom of the tree view. The right panel, 'Frameset View', shows 'Roleset Information' for the frame 'find.01'. It includes fields for ID (find.01), Name (discovery), Arg1 (finder), Arg2 (thing found), and Arg3 (benefactive, found for). Below this is the 'Argument View' section, which contains a grid of buttons for roles 0 through 3.

1.2. Specifications

The specifications for the Baby SRL project follow the Penn Treebank II guidelines (Bies et al. 1995) and the Propbank 3.2 guidelines (Bonial et al. 2013) as closely as possible, with uses from Warner et al. 2012 added.

Due to the nature of adult-child interactions, some revisions and additional specifications are needed.

2. Revisions and Addenda to Penn Treebank II Guidelines

In general, the Penn Treebank II guidelines (Bies et al. 1995) are used, with the addenda given below.

2.1. Imperatives

We neither flag a sentence as an imperative (see Taylor 1996) nor include a null subject.

This screenshot shows the Jubilee 3.26 software interface again. The left panel, 'Treebank View', displays a tree diagram for a sentence from adam01.task. The root node is 'TOP S1'. Below it is a 'VP' node, which contains a 'VB leave 0:0-rel' node, an 'NP 1:1-ARG1 PRP them' node, a 'PP 2:1-ARGM-LOC IN in' node, and an 'ADVP RB there' node. A large block of annotations is shown at the bottom of the tree view. The right panel, 'Frameset View', shows 'Roleset Information' for the frame 'leave.02'. It includes fields for ID (leave.02), Name (give), Arg1 (giver / leaver), Arg2 (thing given), and Arg3 (benefactive / given-to). Below this is the 'Argument View' section, which contains a grid of buttons for roles 0 through 3.

There is no first argument labeled in imperative cases.

2.2. Left-Edge Deletion

We treat utterances with various forms of left edge deletion as sentences, provided that there is a predicate, including gerunds. Examples include sentences that are missing subjects but are not imperatives, as well as any left-edge deletion that sounds like typical adult speech.

An example of typical left-edge deletion is below:

Examples:

The screenshot shows the Jubilee 3.26 software interface with three main panels:

- Treebank View:** Displays a syntactic tree for a sentence. The root node is TOP S1, which contains a S node (with a child VP) and a comment "VBZ looks 0:0-rel". The VP node contains a PP node (with a child IN node labeled "like") and an NP node (with a child NN node labeled "Chatter"). Below the tree is a transcript of child utterances, starting with "## 636 *CHI: hmm ?" and ending with "## 651 *CHI: camera .".
- Frameset View:** Shows Roleset Information for the verb "look". The roleset ID is look.02, with Name : seeming, Arg0 : seemer, Arg1 : seemed like, attribute of arg0, and Arg2 : seemed to.
- Argument View:** A grid for labeling arguments. It has four columns labeled 0, 1, 2, and 3. Rows represent different argument types: M-ADV (V), M-CAU (C), M-DIR (D), M-DIS (I); M-DSP (B), M-EXT (E), M-LOC (L), M-MNR (M); M-MOD (O), M-NEG (N), M-SLC (S), M-PRD (T); M-PRP (P), M-RCL (R), M-REC (R), M-TMP (T); and -UNDEF (U), ERASE (-).

2.3. Fragments

Our definition for fragments is more general than that of Albright et al. (2013). They reserve FRAG for utterances that have component parts that do not join to form a typical syntactic constituent.

"FRAG is used for material that cannot be analyzed as forming a standard syntactic phrase like S or NP. FRAG is commonly used to join a full sentence with the preceding heading. It can also be used to link constituents that are related, but not be a standard syntactic relationship such as SBJ/PRD, coordination, or apposition." (Albright et al. (2013:925))

Because we are working with parent-child interactions, we do not want to over-predict (especially in child utterances) what the intended syntactic structure of an utterance is (see Culicover 1970). Therefore, only known constituent structure is represented in samples labeled as FRAG. Component parts that do not form typical constituents appear as sister nodes under the FRAG node.

Examples:

There is no governor labeled in the preposition SRL.

In the fragment, "cup on table", we do not assume that the child meant either "put the cup on the table" or "the cup is on the table". Instead, we label the clear constituents (e.g., "on table" as a prepositional phrase), but we do not make any unnecessary judgments about the structure that is not

clear. The NP and PP above are coded as sisters. We use the FRAG label for adult and child utterances.

Some challenging cases are cases that have a deictic referent and cases that have a referent in the discourse surrounding the utterance. Each of these are licit utterances in adult language, but are identified as fragments.

In general, if the linguistic expressions do not form expected syntactic constituents then no SRL labels are given. Where 'expected syntactic constituents' refers to the constituency structure expected for that sense.

2.3.1. Prepositional Phrases and Noun Phrases

When a parsed and labeled sample only contains a prepositional phrase or only contains a noun phrase that contains a prepositional phrase, we allow the label FRAG at the top node, if that is the label

The screenshot shows the Jubilee 3.26 software interface with the following details:

- Treebank View:** Displays the parse tree. The root node is labeled "TOP S1". It branches into "FRAG" (RB not) and "PP" (IN in 1:0-rel, NP 2:1-ARGP-OBJ (PRPS your NN mouth)).
- Frameset View:** Shows Roleset Information for the verb "in". ID: in.Destination, Name: Destination, Arg: governor, Arg: object.
- Argument View:** A grid of buttons for selecting arguments, with rows for M-A..., M-C..., M-DI..., M-DL..., M-E..., M-L..., M-M..., M-N..., M-SL..., M-P..., M-R..., M-R..., M-T..., and -UN... ERA... .
- Text Area:** Displays the following transcript entries:
 - ## 83 ^CHI: who (th)at # Daddy ?
 - ## 84 ^CHI: who's [=? who] (th)at ?
 - ## 85 ^CHI: two # three .
 - 86 ^URS: can you put them together # Adam ?
 - ## 87 ^CHI: big horn .
 - 88 ^MOT: not in your mouth .** (This line is highlighted in yellow)
 - 89 ^MOT: let me put them together .
 - ## 90 ^MOT: no # put floor .
 - 91 ^URS: you'll have to pick them up .
 - 92 ^URS: take two # all right ?
 - 93 ^CHI: take two .
 - ## 94 ^CHI: who (th)at ?
 - 95 ^CHI: put record .
 - 96 ^CHI: put record ?
 - ## 97 ^CHI: Adam horn .
 - ## 98 ^CHI: Adam horn # hmm hmm ?

that the automatic parse provides. The Penn Treebank Guidelines do not label such samples with FRAG but rather label the node as PP or NP.

An example is given in the screenshot above for the sentence 'not in your mouth'.

2.4. Gerunds

Gerunds in noun phrases, like 'drinking cowboy' were deleted in this round. However, they were not deleted in the previous round, so, if they were in an BabySRL example, we allowed them to remain.

2.5. Tag Questions

The verbs in tag questions are not given a predicate sense. They are auxiliaries.

2.6. Vocatives and Interjections.

Sentence-initial vocatives and interjections are separated from the sentence as an adjunct to the

sentence node as shown below:

The screenshot shows the Jubilee 3.26 software interface. On the left, the 'Treebank View' displays a parse tree for sentence 396. The root node is TOP S1, which contains an NP (NP Mommy) and an SBAR (WHNP 2:1-ARG2). The WHNP node contains a WP (what) and an S node. The S node contains a VP node with an AUX 's 3:0-rel' and an NP node (NP 4:1-ARG1 DT that). Below the tree, a list of utterances is shown, starting with 390 MOT: "what kind di(o) you have ?". The entry for 396 CHI: "Mommy # what's dat ?" is highlighted in yellow. On the right, the 'Frameset View' shows information for the verb 'be'. It includes fields for ID (be.01), Name (copula), Arg1 (topic), and Arg2 (comment). Below it is the 'Argument View' section, which contains a grid of buttons for various argument types, such as P-GOV (X), P-OBJ (Y), M-ADV (V), etc.

The use of Mommy in Adam23:396*CHI is not an argument of the main predicate, rather, it is a vocative use.

An example of an interjection is shown below:

This screenshot shows the Jubilee 3.26 software interface again. The 'Treebank View' displays a parse tree for sentence 386. The root node is TOP S1, which contains an INTJ node (UH oh) and an S node. The S node contains an NP node (NP 2:1-ARG0 PRP I) and a VP node. The VP node contains an AUX node (AUX have 3:0-rel) and an NP node (NP 4:1-ARG1 DT some NN meat). Below the tree, a list of utterances is shown, starting with 379 MOT: "I like hamburger .". The entry for 386 MOT: "oh # i have some meat ." is highlighted in yellow. On the right, the 'Frameset View' shows information for the verb 'have'. It includes fields for ID (have.03), Name (own, possess), Arg0 (owner), and Arg1 (possession). Below it is the 'Argument View' section, identical to the one in the previous screenshot.

2.7. Pauses.

Pauses, marked by '#' in the data, are transformed into commas. If the pause separates naturally

occurring arguments. Sometimes they separate vocatives and interjections from the sentence, and other times they separate constituents.

When the comma represents a pause between constituents, the constituent structure of the sentence is retained, as shown below:

The screenshot shows the Jubilee 3.26 software interface for editing XML annotations. The main window is divided into three main panels:

- Treebank View:** Displays the constituent structure of a sentence. The root node is **S**, which contains **NP 0:1-ARG0** (PRP I), **VP** (VBG going), and **VP** (VB give 2:0-rel, NP 3:1-ARG2 (PRP you)). Below this, there are additional nodes like **NP 5:1-ARG1** (DT some, NN tea). A scrollable list of annotations follows the tree structure.
- Frameset View:** Shows the roleset information for the verb **give**. It includes fields for ID (give.01), Name (transfer), Arg0 (giver), Arg1 (thing given), and Arg2 (entity given to).
- Argument View:** A grid where semantic roles are assigned to arguments. The grid has columns labeled 0, 1, 2, 3, 4, and 5. Rows represent different semantic role types: M-ADV (V), M-CAU (C), M-DIR (D), M-DIS (I), M-DSP (B), M-EXT (E), M-LOC (L), M-MNR (M), M-MOD (O), M-NEG (N), M-SLC (6), M-PRD (7), M-PRP (P), M-RCL (R), M-REC (9), M-TMP (T), -UNDEF (U), and ERASE (-).

Annotations listed in the Treebank View panel:

```

## 364 *MOT: you what ?
365 *CHI: I cut dem .
366 *MOT: you cut it for me ?
367 *MOT: what kind of meat is this ?
## 368 *CHI: hamburger .
## 369 *MOT: hamburger .
370 *CHI: thank you .
371 *CHI: I am goin(g) give you # some tea .
372 *CHI: I carry some # for me .
373 *MOT: you're carrying some for you .
374 *CHI: I goin(g) eat dem with xxx it's xxx plate .
## 375 *MOT: some what ?
## 376 *CHI: some +...
377 *CHI: I go some for me .
378 *CHI: Mommy # eat it .
379 *MOT: I like hamburger .

```

Buttons at the bottom of the Treebank View panel include [+/-] Add Annotation, [-/+] Remove Annotation, and Edit Tree.

2.8. Uses of 'to be'.

For uses of 'to be', we follow the Penn Tree Bank guidelines, but here we clarify annotations of topic and comment labeling.

Uses of 'to be' that are labeled as having a topic and comment are consistently labeled such that the topic is the ARG1 and the comment is the ARG2. Sometimes, the automatic parser gives a specific ARGM, for example, given a sentence like, 'It is twelve o' clock', the comment 'twelve o' clock' could be labeled as a temporal ARGM. We do not use these ARGM labels. Instead, we replace the label with ARG2.

When the sentence is inverted, the semantic role labels are sequenced according to their position in a non-inverted, declarative sentence, as shown below:

The screenshot shows the Jubilee 3.26 software interface. The left pane, 'Treebank View', displays a parse tree for a child utterance. The root node is TOP S1, which contains an NP (NP Mommy) and an SBAR (WHNP 2:1-ARG2 WP what). The SBAR node contains an S (AUX 's 3:0-rel NP 4:1-ARG1 DT that) and a question mark ?. Below the tree, a transcript of the child's speech is shown, including lines like "what kind do you have?", "I goin(g) see # whoops", and "Mommy # what's dat?". The right pane, 'Frameset View', shows information for the verb 'be'. It includes 'Roleset Information' (ID: be.01, Name: copula, Arg1: topic, Arg2: comment), an empty 'Example' field, and an 'Argument View' grid for M-ADV through M-TMP categories.

The only forms of 'to be' which are not labeled as topic and comment readings are existential uses where a pleonastic 'there' is present, as shown below:

This screenshot shows another instance of the Jubilee 3.26 software interface. The 'Treebank View' pane displays a parse tree for a child utterance. The root node is TOP S1, which contains an INTJ (UH well) and an S (NP EX there VP). The VP node contains an AUX 's 3:0-rel NP 4:1-ARG1 DT another NN knife. The transcript below shows the child saying "well # there's another knife.". The 'Frameset View' pane on the right shows information for the verb 'be' (ID: be.02, Name: existential, Arg1: thing that is), with an empty 'Example' field and an 'Argument View' grid.

With uses of is with progressive verbs, as in 'he is taking the dog for a walk', only the main verb is given a sense.

2.9. Particle Verbs

In many cases, it is difficult to know whether a use of a verb with a preposition is a preposition without an argument or an example of a particle verb.

When we suspect an example is a particle verb use, first we check the Propbank information in the

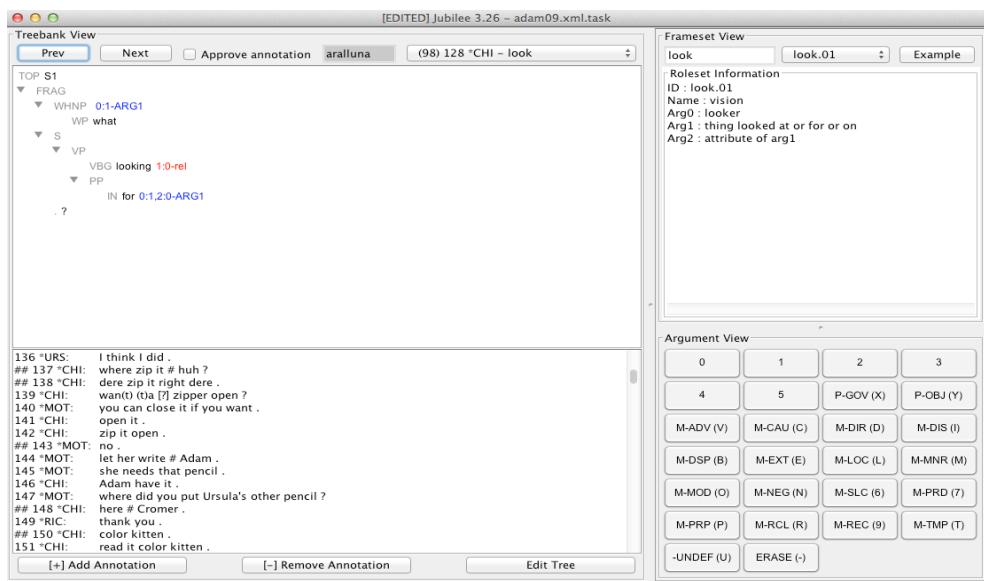
Jubilee window for a sense which lists the particle verb form. For example, if the main verb is 'bring', there is a particle verb form 'bring up' among the choices. If that sense is found, we label the predicate and then the particle using the 'slash' command (under Arguments -> Functions -> /). The other arguments are labeled according to the chosen sense.

The online resources for PropBank are consulted for the relevant sense. They are found online in the Unified Verb Index at <http://verbs.colorado.edu/verb-index/index.php>

If the suspected particle verb is not listed, we check the list of senses added in the last round of Baby SRL annotation http://cogcomp.cs.illinois.edu/page/resource_view/24. If the form is found there, we label the particle as suggested and choose the closest Frameset sense we can. If the suspected particle verb is still missing, we keep the main verb labeled with a 'xx' sense. After consulting other annotators, we may decide to add this to our list of senses. Usually the most generic sense possible that still gives relevant core arguments.

2.10 Stranded Prepositions

For stranded prepositions, we label the form as a continuation. The noun phrase is labeled as the argument role that it fits, and the preposition is then labeled as being part of a continuation using the shift + , command (under Arguments -> Functions -> shift + ,).



2.11. Auxiliaries: Modal Auxiliaries, Do-Support

Uses of modal auxiliaries, such as 'going' in a sentence like 'I'm going to get the book' are sometimes mis-labeled as a main verb by the automatic SRL. Similarly, uses of 'do' in do-support uses, as in 'Do you like this book?' are sometimes labeled as main verbs. In each case, we remove these annotations. We also remove annotations of the verb 'be' when it occurs as an auxiliary in a progressive form, such as 'I am having fun.' or 'Mother is making cake.' In each case, the main verb is annotated with semantic role labels, but the auxiliary is not.

Occasionally, the previous BabySRL annotation has correctly labeled an auxiliary and it has the proper Propbank sense and no semantic roles, as expected for an auxiliary entry. In these cases, we follow the general guideline to not change things that do not need to be changed. We leave the annotations as they

are.

2.12. Use of the Context Window for Disambiguation

When the child says, 'Sit there.', we don't want to assume a missing subject and predicate as in '[You] sit there.', because the utterance could mean a number of things. It could be self narration, as in 'I'm going to sit over there' or a description of the room in which the child is in, 'The big drum is sitting over there.' Each of these uses, however, has the same constituent structure.

When unsure of the constituent structure, the discourse context appearing in the context window can be used to see if there is an indication of the use. For example, if the child is saying, 'I going to get a chair.' then says 'I going', the use of 'going' is likely to be an auxiliary use (which we delete), not a main verb use.

The example below illustrates how the context window can help. The utterance is 'Adam sit down book', which sounds, out of context, as if Adam is setting a book down, however, three lines later, in the context window, his mother says, 'No, you can't sit on my book'. So Adam appears to be self-narrating his intention to sit on the book. The original parse is shown below, followed by the correction inspired by the context:

The screenshot shows the Jubilee 3.26 software interface with three main panes:

- Treebank View:** Displays a parse tree for the sentence "Adam sit down book". The root node is S1. The tree structure is as follows:
 - S1 (TOP)
 - S (TOP)
 - NP (0:1-ARG1)
 - NNP Adam
 - VP
 - VBP sit 1:0-rel
 - PRT
 - RP down 1:0,2:0-rel
 - NP
 - NN book
- Frameset View:** Shows the frame set for the verb "sit".
 - ID: sit.01
 - Name: be in a position
 - Arg1: thing sitting
 - Arg2: location or position
- Argument View:** A grid of buttons for selecting arguments:

0	1	2	3
4	5	P-GOV (X)	P-OBJ (Y)
M-ADV (V)	M-CAU (C)	M-DIR (D)	M-DIS (I)
M-DSP (B)	M-EXT (E)	M-LOC (L)	M-MNR (M)
M-MOD (O)	M-NEG (N)	M-SLC (6)	M-PRD (7)
M-PRP (P)	M-RCL (R)	M-REC (9)	M-TMP (T)
-UNDEF (U)	ERASE (-)		

Below the Treebank View pane, there is a text area containing a sequence of numbered lines (## 53 to ## 68) representing child utterances. Lines 59 and 60 specifically mention "Adam sit down book" and "no.", respectively.

At the bottom of the interface are three buttons: "[+] Add Annotation", "[-] Remove Annotation", and "Edit Tree".

[EDITED] Jubilee 3.26 – adam02.xml.task

Treebank View

Prev Next Approve annotation gold (39) 59 *CHI – sit

```

GOLD   0:1-ARG1 1:0,2:0-rel 3:1-ARG2(S1 (S (NP (NNP Adam)) (VP (VBP sit) (PRT (RP down))) (NP (NN book)))
aralluna 0:1-ARG1 1:0,2:0-rel  (S1 (S (NP (NNP Adam)) (VP (VBP sit) (PRT (RP down))) (NP (NN book))) (. .)
auto    0:1-ARG1 1:0,2:1-rel  (S1 (S (NP (NNP Adam)) (VP (VBP sit) (PRT (RP down))) (NP (NN book))) (. .)

```

TOP S1

- ▼ S
 - ▼ NP 0:1-ARG1
 - NNP Adam
 - ▼ VP
 - VBP sit 1:0-rel
 - ▼ PRT
 - RP down 1:0,2:0-rel
 - ▼ NP 3:1-ARG2
 - NN book

..

55 *CHI: train track ?
56 *CHI: train track ?.
57 *CHI: read dat train track .
58 *CHI: train train .
59 *CHI: Adam sit down book .
60 *CHI: no .
61 *CHI: Adam sit .
62 *MOT: no # you can't sit on my book .
63 *CHI: choo+choo train .
64 *MOT: what kind is that ?
65 *CHI: choo+choo train .
66 *MOT: what is that ?
67 *CHI: train track .
68 *MOT: this is a diesel locomotive .
69 *MOT: can you say that ?
70 *CHI: diesel motor [?] no [?] no [?] moto(r) [?].

[+] Add Annotation [-] Remove Annotation Edit Tree

Frameset View

sit sit.01 Example

Roleset Information

ID : sit.01
Name : be in a position
Arg1 : thing sitting
Arg2 : location or position

Argument View

0	1	2	3
4	5	P-GOV (X)	P-OBJ (Y)
M-ADV (V)	M-CAU (C)	M-DIR (D)	M-DIS (I)
M-DSP (B)	M-EXT (E)	M-LOC (L)	M-MNR (M)
M-MOD (O)	M-NEG (N)	M-SLC (6)	M-PRD (7)
M-PRP (P)	M-RCL (R)	M-REC (9)	M-TMP (T)
-UNDEF (U)	ERASE (-)		

2.13. Embedding Verbs with Infinitival Complements in Child Utterances

In some of the child utterances, embedding verbs which take complements containing infinitive verbs are confusing for the parser. When the child is saying what an adult might say as, 'I am going to pick up', he or she says, 'I going to pick up'. Similarly, 'Do you want to eat one?' might come out in a child utterance as 'Do you want eat one?'

In such cases, the parser often does not put an S node after the embedding verb. However, in the case of a verb like 'want', the S node is the parent node of the semantic role of the thing wished, so we add the S node in order to have the proper tree structure for the semantic role label. An example of a parent utterance with 'want' and a child utterance with 'want' are shown below.

The screenshot shows the Jubilee 3.26 interface with two main panes. The left pane, 'Treebank View', displays a parse tree for the sentence 'PRP you VB want 1:0-rel S 2:2-ARG1 VP TO to VB look NP PRPS your NN bag'. Below the tree is a list of annotations from line 8 to 46. Annotations 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, and 46 are highlighted in yellow. Buttons for '[+ Add Annotation]', '[- Remove Annotation]', and 'Edit Tree' are at the bottom. The right pane, 'Frameset View', shows a 'Roleset Information' table for the verb 'want' with fields: ID: want.01, Name: want, desire, Arg0: Wanter, Arg1: thing wanted, Arg2: beneficiary, Arg3: in-exchange-for, Arg4: from. Below it is an 'Argument View' grid with four rows and four columns of buttons labeled 0, 1, 2, 3, 4, 5, P-GOV (X), P-OBJ (Y), M-ADV (V), M-CAU (C), M-DIR (D), M-DIS (I), M-DSP (B), M-EXT (E), M-LOC (L), M-MNR (M), M-MOD (O), M-NEG (N), M-SLC (6), M-PRD (7), M-PRP (P), M-RCL (R), M-REC (9), M-TMP (T), -UNDEF (U), ERASE (-), M-MOD (O), M-NEG (N), M-SLC (6), M-PRD (7), M-PRP (P), M-RCL (R), M-REC (9), M-TMP (T), -UNDEF (U), and ERASE (-).

2.14. 'Thank you' and 'Please'

Occasionally, the BabySRL parser has labeled 'thank' as a verb in the expression 'Thank you'. If it is tagged correctly, with 'you' as the person thanked, we leave the annotation as it is. If it is tagged incorrectly, we remove the annotation. If it is not tagged, we do not add a tag.

'Please' is treated as an interjection and does not get a verb sense label

2.15. Co-indexing of Nouns

There are two cases in which we co-index nouns. One case is when there is a continuation that is split, and the other is when there is a complex noun phrase.

[EDITED] Jubilee 3.24 – adam06.xml.task

File View Treebank Frameset Argument Help

Treebank View

Prev Next (256) 421 *URS – be auto

TOP S1

- ↳ SBARQ
 - ↳ WHNP 0:1-ARG1
 - ↳ WP what
 - ↳ SQ
 - ↳ AUX 's 1:0-rel
 - ↳ NP 0:1:2:1-ARG1
 - ↳ DT the
 - ↳ NN doggie
 - ↳ VP
 - ↳ VBG doing
 - ↳ PP
 - ↳ TO to
 - ↳ NP
 - ↳ DT the
 - ↳ NN cowboy
 - ↳ NN boot

?

414 *CHI: go dere ?
 ## 415 *CHI: oh # doggie .
 ## 416 *CHI: nineteen twelve .
 ## 417 *CHI: doggie doggie doggie .
 418 *CHI: stop it .
 419 *CHI: stop it # cowboy boot .
 420 *CHI: stop it .
421 *URS: what's the doggie doing to the cowboy boot ?
 422 *CHI: doggie doggie doing doggie .
 423 *CHI: read dat .
 424 *CHI: read dat cowboy book .
 425 *CHI: read dat # Mommy .
 ## 426 *CHI: nineteen twelve .
 427 *CHI: read dat # Mommy .
 ## 428 *CHI: dat one .
 ## 429 *MOT: oh # that one .

[+] Add Annotation [-] Remove Annotation Edit Tree

Frameset View

be be.02 Example

Roleset Information

ID : be.02
 Name : existential
 Arg1 : thing is

Argument View

0	1	2	3
4	5	P-GOV (X)	P-OBJ (Y)
M-ADV (V)	M-CAU (C)	M-DIR (D)	M-DIS (I)
M-DSP (B)	M-EXT (E)	M-LOC (L)	M-MNR (M)
M-MOD (O)	M-NEG (N)	M-SLC (6)	M-PRD (7)
M-PRP (P)	M-RCL (R)	M-REC (9)	M-TMP (T)
-UNDEF (U)	ERASE (-)		

Co-indexing of nouns is done in Jubilee by labeling one noun according to its semantic role and then using the up-arrow comma key to co-index the other one.

2.15. Other Senses

There are occasionally uses which do not have a sense in the Frameset View in the Jubilee annotation tool, but which have a sense in the online Unified Verb Index. When this is the case, we label the use as 'verbXX' and assign the semantic roles as they are described for the given use in the Unified Verb Index.

2.16. Redundancy in Child Utterances

In some cases of child utterances, there are strings of redundant words or phrases such as CHI: 'Jump, jump, jump, ...' In such cases, only the first instance of a word or phrase is annotated while subsequent uses of the word or phrase are left unedited and their corresponding annotations removed.

Note that this only applies when the word or phrase is repeated in a continuous stream; if there are breaks in the speech, such as CHI: 'Jump.' MOT: 'What?' CHI: 'Jump' or CHI: 'Jump, Mommy, Jump', each instance is annotated.

1250 *CHI:take it off.
 1251 *CHI:take it off.
 1252 *CHI:take it off.
 1253 *URS:can I help you # Adam ?
 ## 1254 *CHI:upsadaisy .
 ## 1255 *CHI:xxx .
 1256 *CHI:come down down down .
 ## 1257 *CHI:blow that man down .
 ## 1258 *CHI:blow man down .
 ## 1259 *CHI:blow man down .
 1260 *URS:now let's make a little one for a head .
 1261 *CHI:I have enough .
 1262 *CHI:just like Mommy with two little ears .

3. Revisions and Addenda to Part of Speech Tagging

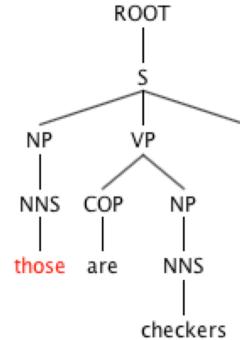
The part of speech tagging, as described in Santorini 1990, is used for terminal and non-terminal nodes, with some modifications.

3.1. Distinctions Between Determiner and Noun Phrase Uses

When a word which can be a determiner, such as 'this' and 'those' occurs in a way that it is co-referential with a noun phrase and behaves syntactically as a noun phrase, we label it as a noun phrase, marking whether it is singular (NN) or plural (NNS). This instruction only applies when the noun phrase takes an argument label.

In the sentence 'those are checkers', 'those' is labeled an NNS. This method is used by Pearl & Sprouse 2013a, 2013b in the CHILDES data as well (shown below).

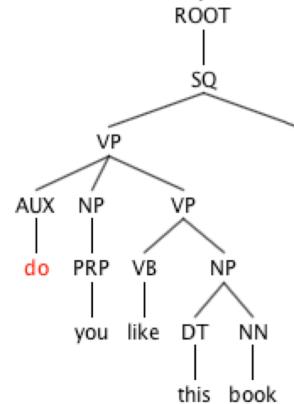
From file: /Applications/stanford-tregex-2014-10-26/examples/CHILDESTreebank-curr-2/brown-adam.parsed



those are checkers .

When 'this' is used as a determiner as shown in the sentence 'Do you like this book?', below, it is labeled as a determiner.

From file: /Applications/stanford-tregex-2014-10-26/examples/CHILDESTreebank-curr-2/brown-adam.parsed



do you like this book .

This decision was made, in contrast to Santorini 1990, in order to reflect the syntactic behaviour more accurately and as something which could, in the future, facilitate discourse resolution of co-referential expressions.

4. Revisions and Addenda to PropBank Guidelines 3.2

4.1. Semantic Role Labels for Prepositions

The most significant addendum to the PropBank Guidelines 3.2 (Bonial et al. 2013) is the addition of semantic role labeling for prepositions, (Srikumar & Roth 2013).

A second topic is what to do with examples that seem to stretch the senses of existing senses in PropBank. Many are examples of phrasal verbs. We make a decision to the best of our abilities regarding which existing sense they most closely resemble. Several decisions are discussed in the spreadsheet of annotator decisions, which is in this repository.

If it is necessary to create a new sense for a predicate, we keep a file of the examples and justification. If it is agreed to be a new sense, we save the documentation and request to have it added to official PropBank categories.

5. Known Charniak Parser Errors and Semantic Role Labeler Errors

There are several known errors with the semantic role labeler and Charniak Parser for which we have guidelines.

5.1. Faux Predicates

One common error is when linguistic expressions that can be used as either predicates or nominals are interpreted as predicates. For example, when the child says, 'Big drum.', the parser takes 'drum' to be a predicate. As a result, the SRL labels the predicate according to its most likely sense. In cases like these, we remove the verb annotation from the word 'drum' and remove any additional semantic role labels since there are no semantic roles of a predicate that we can discern based on the noun phrase.

If the SRL detects a nominal as a predicate, but the sentence contains an actual predicate, we move the relation annotation to the correct predicate and label the arguments that are present accordingly.

5.2. Known Errors in the Automated Output

5.2.1. Verb and Preposition Confusion

The confusion of verbs and prepositions is attested with the linguistic expression 'like', which can be either. When this happens, the sense must be changed to reflect the correct reading.

5.2.2. Verb Sense Confusion

The pipeline SRL tends to label 'fell', the past of 'fall' as the verb 'to fell', and 'putting' as derived from 'to putt' rather than 'to put'. These automated errors are corrected.

5.2.3. Missing Copula Instances

When the copula verb occurs as a contraction or in the third person form, the parser often misses the instance. Annotators added an example in this case.

6. Additional Guidelines

6.1. SRL Scope

The scope of SRL labels differs for predicates and for prepositions.

The main relation expressed by a predicate is labeled directly on the leaf node of the predicate. The arguments of the predicate receive a semantic role label on the lowest projection which contains all parts of the argument, but not on leaf nodes. For example, if the argument of a verb contains a noun

and a determiner, such as 'the man', then the SRL goes on the NP that has 'the' and 'man' as daughters. If the NP consists of a single leaf, for example, if the argument is a proper name, like 'John', then the label still goes on the non-terminal node which dominates only the leaf node containing 'John'.

The main relation expressed by a preposition is labeled directly on the leaf node of the preposition. Each preposition has a governor and an argument.

6.2. Adjudication with Pearl & Sprouse Annotations

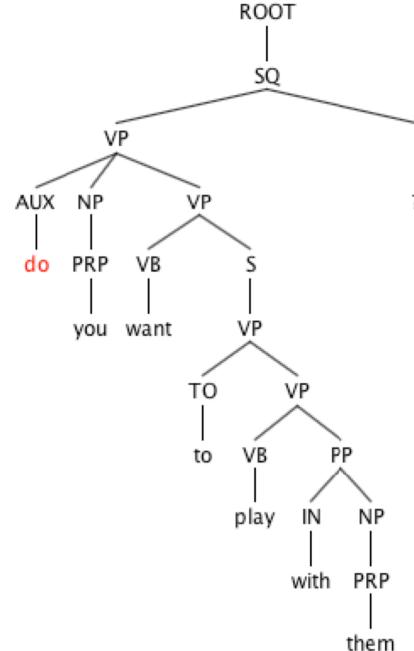
6.2.1. AUX and COP Labels

The Pearl & Sprouse annotations use 'COP' as a node for copular verbs. We use AUX for copular verbs and for auxiliaries. If a main use of a verb is mis-labeled as an AUX, we leave the label as it is, following the general guideline to change the parse as little as possible.

6.2.2. VP Insertion over COP and AUX

The Pearl & Sprouse annotations insert a VP node over copular and auxiliary nodes in cases such as do-support:

From file: /Applications/stanford-tregex-2014-10-26/examples/CHILDESTreebank-curr-2/brown-adam.parsed



do you want to play with them ?

The Charniak Parser will not insert the VP above AUX, but will leave AUX dominated by the SQ node as a sister to the NP and the VP. The Charniak parse is what we want, we do not change the parse to match the Pearl & Sprouse parser decision.

References

- Albright, Daniel, Arrick Lanfranchi, Anwen Fredriksen, William Styler, Collin Warner, Jena Hwang, Jinho Choi, Dmitriy Dligach, Rodney Nielsen, James Martin, Wayne Ward, Martha Palmer, Guergana Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of American Medical Informatics Association*. Vol 0:1-9.
- Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. *Bracketing Guidelines for the Treebank II-style Penn Treebank Project*. University of Pennsylvania.
- Bonial, Claire, Julia Bonn, Kathryn Conger, Jena Hwang, and Martha Palmer. 2014. PropBank Semantics of New Predicate Types. *The 9th Edition of Language Resources and Evaluation Conference, Reykjavik, Iceland*.
- Charniak, Eugene and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.
- Culicover, Peter. 1970. One more can of beer. *Linguistic Inquiry* 1, 366-369.
- Pearl, L. & Sprouse, J. 2013a. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20, 23-68.
- Pearl, L. & Sprouse, J. 2013b. Computational Models of Acquisition for Islands. In J. Sprouse & N. Hornstein (eds), *Experimental Syntax and Islands Effects*. Cambridge University Press, 109-131.
- Punyakanok and D. Roth and W. Yih. 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, 34 Vol 2
- Santorini, Beatrice. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd Printing)*.
- Srikumar, Vivek and Dan Roth. 2013. Modeling semantic relations expressed by propositions. *Transactions of the Association for Computational Linguistics*, 1:231-242.
- Taylor, Ann. 1996. *Bracketing Switchboard: An Addendum to the Treebank II Guidelines*.
- Warner, Colin, Arrick Lanfranchi, Tim O'Gorman, Amanda Howard, Kevin Gould, Michael Regan. 2012. *Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines*. Institute of Cognitive Science, University of Colorado at Boulder.