Gold Standard Annotations for Preposition and Verb Sense with Semantic Role Labels in Adult-Child Interactions

Lori Moon
University of Illinois at
Urbana-Champaign
aralluna@illinois.edu

Christos Christodoulopoulos
Amazon Research
chrchrs@amazon.co.uk

Cynthia Fisher
University of Illinois at
Urbana-Champaign
clfishe@illinois.edu

Sandra Franco
Intelligent Medical Objects
Northbrook, IL USA
sfranco@imo-online.com

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

This paper describes the augmentation of an existing corpus of child-directed speech. The resulting corpus is a gold-standard labeled corpus for supervised learning of semantic role labels in adult-child dialogues. Semantic role labeling (SRL) models assign semantic roles to sentence constituents, thus indicating who has done what to whom (and in what way). The current corpus is derived from the Adam files in the Brown corpus (Brown, 1973) of the CHILDES corpora, and augments the partial annotation described in Connor et al. (2010). It provides labels for both semantic arguments of verbs and semantic arguments of prepositions. The semantic role labels and senses of verbs follow Propbank guidelines (Kingsbury and Palmer, 2002; Gildea and Palmer, 2002; Palmer et al., 2005) and those for prepositions follow Srikumar and Roth (2011). The corpus was annotated by two annotators. Inter-annotator agreement is given separately for prepositions and verbs, and for adult speech and child speech. Overall, across child and adult samples, including verbs and prepositions, the κ score for sense is 72.6, for the number of semantic-role-bearing arguments, the κ score is 77.4, for identical semantic role labels on a given argument, the κ score is 91.1, for the span of semantic role labels, and the κ for agreement is 93.9. The sense and number of arguments was often open to multiple interpretations in child speech, due to the rapidly changing discourse and omission of constituents in production. Annotators used a discourse context window of ten sentences before and ten sentences after the target utterance to determine the annotation labels. The derived corpus is available for use in CHAT (MacWhinney, 2000) and XML format.

1 Introduction

The study of human language acquisition has greatly benefited from the availability of corpora of language use to, by, and around young children. The CHILDES project (MacWhinney, 2000) makes available transcribed corpora of adult-child dialogue in English and in a growing set of other languages.[1] In recent years, annotations have been added to some CHILDES corpora, including part-of-speech tagging, syntactic parsing, and the identification of grammatical roles (Pearl and Sprouse, 2013; Sagae et al., 2010; Sagae et al., 2007).

In the present paper, we describe an ongoing project that adds a new layer of annotation to selected CHILDES corpora, a hand-checked corpus of semantic role labels that provides a shallow semantic analysis of sentences' predicate-argument structure. Our goal is to support the development of computational models of language acquisition that explore how children come to interpret sentences, assigning semantic roles to sentence constituents to determine who does what to whom in each sentence. An additional goal is to provide new resources for testing the ability of trained NLP systems to generalize to new domains, in this case the challenging linguistic environment of dialogues between toddlers, who are in incomplete stages of language acquisition, and adults.

[1] Corpora and documentation are available at https://childes.talkbank.org.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3004–3014
Santa Fe, New Mexico, USA, August 20-26, 2018.

Semantic role labeling (SRL) is a common task in NLP. For each predicate, an SRL system identifies sentence constituents and assigns to them argument (e.g., agent, patient) or adjunct (e.g., locative, manner) roles. Usually SRL refers to labeling verb semantic roles, but it has been extended to nominal predicates (Meyers et al., 2004), as well as prepositions (Srikumar and Roth, 2011; Srikumar, 2013; Schneider, 2016). SRL has proven useful in areas such as question answering and textual entailment. Annotated data sets for training and evaluating the performance of SRL systems are time-consuming to construct, but new types of annotated data are important for modeling early language acquisition, and for testing the ability of SRL systems to generalize across varieties of language use.

The corpus described in this paper augments portions of an existing partial annotation of child-directed speech corpora, as described in Connor et al. (2010), with additional verb semantic role labeling and preposition semantic role labeling of the adult and child speech. This annotation project involved the Brown corpus (Brown, 1973) from the CHILDES corpora, a classic study of interactions between young children and their caregivers and other adults. The Brown corpus contains data collected in natural conversational contexts at home, and includes multiple sessions with each of three children, called 'Adam', 'Eve', and 'Sarah', learning English as their first language. The transcribed corpora are freely available on the CHILDES site, and have been involved in many analyses of the input for first language acquisition. The Brown corpus was chosen for this project in part because it has been the focus of some of the morphosyntactic annotations mentioned above. The already-released partial annotation of the Brown corpus included verb semantic role labels for some of the parental speech in the Adam, Eve, and Sarah corpora, as described below. In the present extension, we annotate all speech containing a preposition or a verb, spoken by adults or by the child, in Adam files 01-23 for verb or preposition sense and semantic role labels. This project used trained annotators, who followed existing specifications for semantic-role annotation, and where necessary, developed additional specifications for the task.

The previously released partial annotation of child-directed speech was used to train and evaluate a 'BabySRL' model that learned to interpret sentences based on simple representations of syntactic structure, derived from a constrained distributional analysis of child-directed speech, amplified by simple built-in expectations about predicate-argument structure (Connor et al., 2010; Connor, 2011; Connor et al., 2013). This system demonstrated that simple syntactic features based on the set of nouns in a sentence can guide early steps in language learning. The BabySRL learned to label the first of two nouns as an agent in simple sentences with invented verbs (e.g., *'Adam krads Mommy'*), replicating the linguistic behavior of toddlers, as shown in previous experimental work with children (Gertner et al., 2006). The model also made striking errors with two-noun intransitive sentences with invented verbs (e.g., *'Adam and Mommy krad'*), as do toddlers learning English (Gertner and Fisher, 2012). These errors diminished as the model learned to find verbs, gaining data about the importance of verb position in English sentences. These results showed that partial sentence structures grounded in sets of nouns are useful for learning in natural corpora; the model's constrained distributional learning component offers one account of where these partial sentence structures might come from during early acquisition. It also appeals to powerful evidence that infants detect distributional cues that are relevant to discovering grammatical categories (e.g., (Lany and Saffran, 2010; Shi and Melancon, 2010)).

In addition to supporting models of human language development, annotated corpora of adult-child dialogue can provide a useful context in which to evaluate the robustness of NLP learning models. Toddlers producing their early word combinations often omit the function words that support high-accuracy part-of-speech tagging, parsing and semantic-role assignment (e.g., Brown, 1973). Despite these omissions, however, they are often understood by their adult interlocutors. Understanding speech with missing elements requires a flexible knowledge of language.

The partially annotated corpus of child-directed speech that supported the development of the BabySRL model had two main limitations. First, it annotated only verbs' arguments. Second, it annotated only parental speech, leaving out the children's own contributions to the conversation. This work complements the previous corpus, and corpora such as that used in Fernald et al. (2009), which model discourse environments with child-directed speech. Our corpus does not have visual information, but it is compatible with semantic role labeling methods that are standardly used with other corpora.

The corpus labels both adult and child speech in the dialogue, adding to the noisiness of the data and providing a realistic model for the speech of young children. Naturally occurring noise in data is an interesting theoretical problem. Testing existing automatic semantic role labelers on these data provides an engineering challenge for improving tools in a new domain.

The corpus is of dialogue, so semantic arguments of a verb can appear across interlocutors. Though preposition semantic role labels are only given at the sentence level and not at the discourse level (Srikumar 2013:7), they can be used in a way that facilitates identifying semantic arguments across sentences. For example, if one person says, *'Where did the toy fall?'*, and the next line of dialogue is *'on the floor'*, this corpus is annotated in such a way that could support recovery of semantic roles across sentences in a dialogue. That is, the first sentence is annotated with *'fall'* as the main verb. The word *'where'* is the location argument of *'fall'*. The next utterance *'on the floor'* is annotated with semantic role labels for the head preposition *'on'* with the sense of *location*.

These gold labels also provide a new domain for training an automated preposition semantic role labeler. The usefulness of labeling prepositions was demonstrated in the earlier example of finding answers to questions in a dialogue. Preposition roles also add structure to verb semantic role labelers because the prepositions are often contained in an argument of the main predicate. With this corpus, if a verb has a semantic role assigned to a prepositional phrase, it follows that the preposition takes the same verb as a governor. This information can be used to tie prepositions to governing verbs occurring in the discourse.

In section 2, we explain semantic roles in more detail, discussing Propbank and the automatic preposition semantic roles of Srikumar and Roth (2011). Section 3 describes the annotation tool that we used along with modifications added to the tool for our project. In section 4 we discuss special problems presented by the CHILDES data set (Brown, 1973), explaining why it serves to meet an existing need in the community and foster more scientific discovery. Section 5 provides a break-down of our IAA measures and what they mean in terms of accuracy of the labels. Section 5.1 describes the ratings on the held-out data and what users can expect in the final version of the corpus. We note that this data set contains inherent noise and show that annotator scores reflect noise in the areas that have below 90κ scores, and achieve scores above 90κ in other areas. Section 6 discusses the availability and licensing of the corpus, and section 7 concludes the paper.

2 Semantic Role Labels

2.1 Verb Semantic Role Labeling with Propbank

Propbank (Gildea and Palmer, 2002; Kingsbury and Palmer, 2002; Palmer et al., 2005) provides resources for labeling semantic roles for verbs. The original Propbank corpus included a large hand-annotated corpus of semantic verb-argument relations, and extensive guidelines for annotating verb semantic roles in new corpora (Bonial et al., forthcoming). Propbank added semantic role labels to sentences parsed according to Penn Treebank Guidelines (Mitchell et al., 1993). Other corpora have been annotated with Propbank verb senses and semantic roles, including discourses and SMS.[2]

For each verb, Propbank lists a set of senses for the verb and the licit semantic arguments for that sense. The list of semantic arguments includes core arguments like the agent and patient of transitive verbs as well as directional and locational phrases that commonly occur with the verb. Propbank annotations involve annotating the span of each of a verb's arguments. For example, for the verb *'put'*, sense number 01, has the licit semantic roles of *Arg0, Arg1,* and *Arg2*, which are *putter*, *thing put*, and *where put*, respectively. It is not the case that all of these are present in every use of the verb *put* in a corpus.[3]

2.2 Preposition Semantic Role Labeling

Labeling preposition semantic roles helps with NLP tasks by providing additional shallow semantic information about prepositions and their semantic relation to other words in a sentence.

[2] For current information on corpora with Propbank annotations, see `https://propbank.github.io`.

[3] Senses are from the index of Propbank and FrameNet (Baker et al., 1998) for English at `http://verbs.colorado.edu/propbank/framesets-english-aliases/`.

The preposition semantic role labeler is the one described in Srikumar and Roth (2011) and Srikumar (2013). This role labeler was used due to the relatively small number of observed preposition semantic roles, and its integration with the CogComp NLP pipeline used in the current project.[4] Each preposition has the potential semantic roles associated with it of GOVERNOR and GOVERNED. The governed argument in the phrase *'to the store'* is *'the store'*. It is generally a noun phrase that follows the preposition.

The governor of the preposition can be a verb that takes a preposition as an argument. For example, in the sentence *'Take the cart to the store'*, the preposition *'to'* has the verb *'take'* as a governor. The governor can also be a noun phrase. For example, in the sentence *'Give me the horse with the blue mane.'* the governor of *'with'* is *'the horse'*.

3 Jubilee Annotation Tool

The annotation tool was based on the Jubilee tool by (Choi et al., 2010)[5] and the modified version is available at `https://gitlab-beta.engr.illinois.edu/babysrl-group/jubilee`. The original annotation tool used the Penn Treebank annotations (Mitchell et al., 1993) and Propbank's framesets,[6] and after an initial automatic SRL annotation phase, it allowed the annotators to modify the predicate sense and assign the associated semantic roles to constituents of the sentence.

We extended the tool in several ways to accommodate the annotation of children's utterances and prepositional SRL, and provided other improvements for the convenience of the annotators. A summary of the changes is as follows: We used a JSON format for the annotation files and stored the syntactic trees internally, instead of relying on a separate treebank; we added the ability to edit syntax trees with bracket highlighting; we added the ability to delete entire annotations if they had no real predicate, and we allowed the creation of new predicates when an entry was missing; we extended the context window to show more dialogue context; we added an inter-annotator agreement (IAA) calculator; we added the ability to view predicate information via a link to the Propbank website;[7] we added a bookmarking ability to allow annotators to save difficult annotation cases for future discussion.

3.1 Data Sources

The BabySRL Corpus (Connor, 2011; Connor et al., 2013) annotated a portion of the Adam, Eve, and Sarah files from the Brown corpus (Brown, 1973) with verb semantic role labels.[8] The project only annotated adult speech and omitted uses of the copula verb *'to be'*. We imported data from this corpus into the tool. The imported data did not include verb senses, but it had verb semantic role labels.

Utterances that were not part of the previous derived corpus were automatically parsed and labeled for verb and preposition semantic roles and verb and preposition sense using the NLP pipeline tools available through the Cognitive Computation Group.[9] Files were preprocessed from CHAT format to JSON format in order to annotate the data in the Jubilee tool.

After the first five Adam files were annotated, there was additional preprocessing on the xml files. Most of this preprocessing was for the purpose of working with transcriptions. There are several conventions used in the transcription that were changed. Adam's use of interdental fricatives, represented orthographically as *'th'* (phonetically [ð] and [θ] in many adult dialects of North American English), were transcribed orthographically in the CHILDES data with *'d'*, resulting in *'dat'* for *'that'*, for example. Because the entire discourse was not phonetically transcribed, we found the use of orthography

[4] An anonymous reviewer mentioned that we should explain why we did not use the more recent senses of Schneider (2016). Although there is more coverage with the preposition SRL of Schneider (2016), it uses fine-grained relations that are rarely observed. Due to the high ambiguity of preposition use in toddler speech, even the small number of distinctions in Srikumar and Roth (2011) required large amounts of discussion and calibration between annotators.

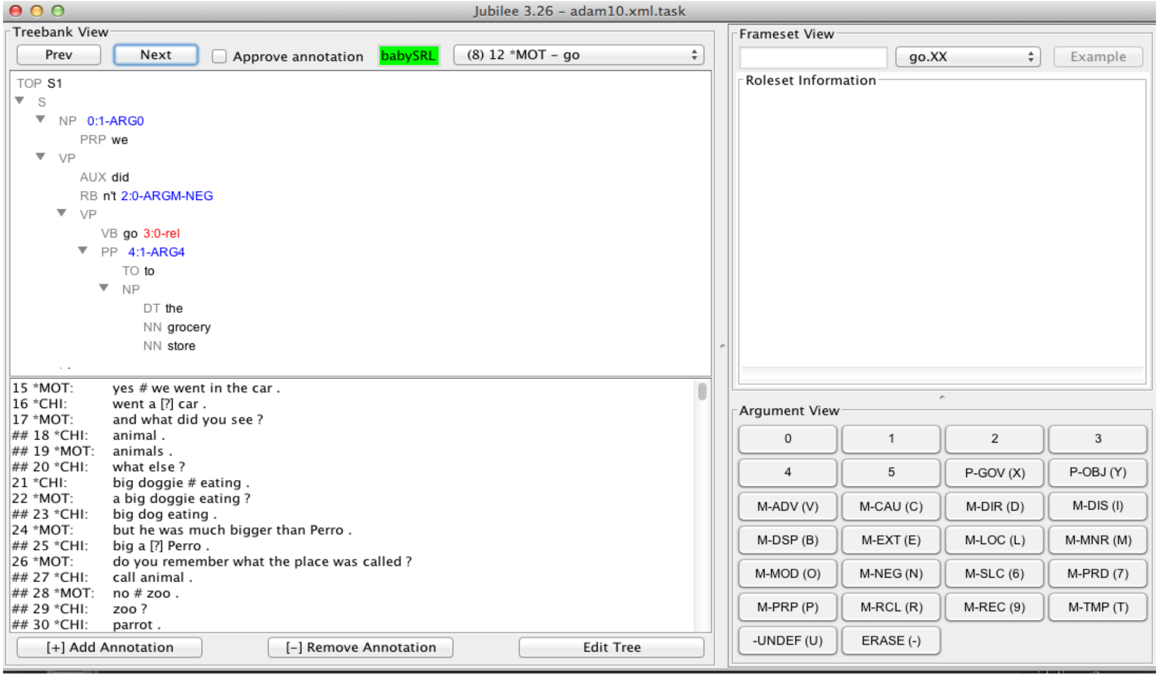
[5] The Jubilee tool is described at `https://code.google.com/archive/p/propbank/`

[6] Framesets contain a verb's sense, the associated semantic roles, and examples of uses of the intended sense in corpora. For examples, see `https://verbs.colorado.edu/propbank/framesets-english/`

[7] `http://verbs.colorado.edu/propbank/framesets-english-aliases/`

[8] The derived corpus is available through TalkBank derived corpora `https://childes.talkbank.org/derived/`

[9] The relevant tools are available at `https://cogcomp.org/page/software/` under 'NLP Tools'. The specific set of tools and instructions for replicating the results are available at `https://gitlab-beta.engr.illinois.edu/babysrl-group/babysrl-corpus`.



```
TOP S1
▼ S
  ▼ NP  0:1-ARG0
      PRP we
  ▼ VP
      AUX did
      RB n't 2:0-ARGM-NEG
    ▼ VP
        VB go 3:0-rel
      ▼ PP  4:1-ARG4
          TO to
        ▼ NP
            DT the
            NN grocery
            NN store
```

```
15 *MOT:      yes # we went in the car .
16 *CHI:      went a [?] car .
17 *MOT:      and what did you see ?
## 18 *CHI:   animal .
## 19 *MOT:   animals .
## 20 *CHI:   what else ?
21 *CHI:      big doggie # eating .
22 *MOT:      a big doggie eating ?
## 23 *CHI:   big dog eating .
24 *MOT:      but he was much bigger than Perro .
## 25 *CHI:   big a [?] Perro .
26 *MOT:      do you remember what the place was called ?
## 27 *CHI:   call animal .
## 28 *MOT:   no # zoo .
## 29 *CHI:   zoo ?
## 30 *CHI:   parrot .
```

Figure 1: Example from the previous derived BabySRL corpus. 'babySRL' is highlighted in green, indicating that the labels come from the previous corpus. The context window at the bottom shows the surrounding discourse.

to represent the pronunciation inconsistent with the way in which the majority of the corpus was transcribed, so we preprocessed these examples to have standard orthography. Unknown words, transcribed as *'xxx'* were removed, as were symbols for spelling out loud, such as *'@c'*, which indicated someone speaking the name of the letter 'c'. All indicated pauses, which used the symbol *'#'* were changed to commas. Characters following underscores, along with the underscores, were removed.[10]

For utterances that were annotated in the previous round, annotators saw the gold-standard annotation without the verb sense listed. An example is given in Figure 1. Near the center-top of the bar in the Jubilee tool, there is a window that says 'babySRL', highlighted in green. Annotators could see from this window whether they were annotating utterances from the previous corpus, the automatic annotation, or viewing utterances that they had already annotated. Below the window containing the parsed and labeled utterance, there is a context window that shows some of the surrounding discourse.

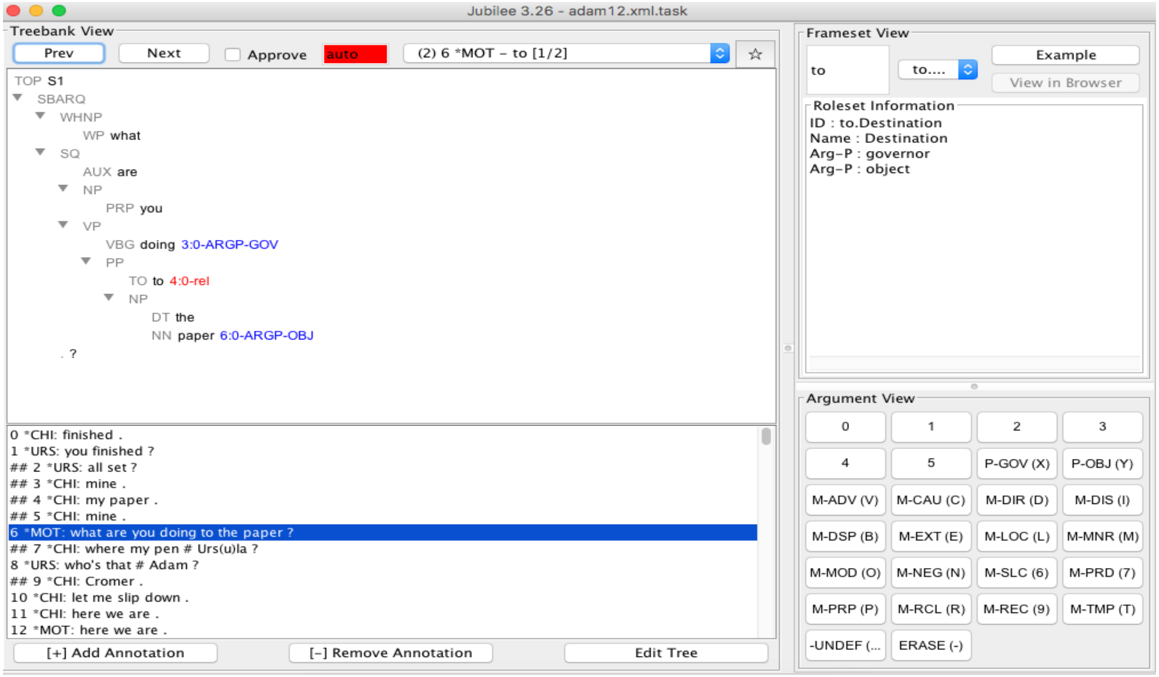
To the right of the main windows, there is a window in which role set information for a sense appears. In the previous BabySRL example, no sense was imported. Annotators choose the appropriate sense. They could use semantic role labels to help determine the intended sense. The 'View in Browser' button takes the annotator directly to the Propbank online index, where annotators can check senses.

For utterances new to this round of annotation, annotators saw the automated parse and semantic role labeling. Figure 2 shows the output of the NLP pipeline, as the annotator would see the data. Rather than a green window labeled 'babySRL' there is a red window labeled 'auto', indicating that the annotator is viewing an automatic parse and SRL. In this example, the tool assigns the most likely sense, and the roles associated with the sense are shown in the Frameset View window.

4 Annotation Guidelines

The syntactic parses follow the Penn Treebank Guidelines (Santorini, 1990; Mitchell et al., 1993) as outlined in Bies et al. (1995) and the modified guidelines in Warner et al. (2012). Annotators were

[10] These changes are in BrownXMLReader.java at `https://gitlab.engr.illinois.edu/babysrl-group/babysrl-corpus/blob/master/src/main/java/edu/illinois/cs/cogcomp/babySRL/corpus/xml/BrownXMLReader.java`



```
TOP S1
▼ SBARQ
  ▼ WHNP
      WP what
  ▼ SQ
      AUX are
    ▼ NP
        PRP you
    ▼ VP
        VBG doing 3:0-ARGP-GOV
      ▼ PP
          TO to 4:0-rel
        ▼ NP
            DT the
            NN paper 6:0-ARGP-OBJ
  . ?
```

```
0 *CHI: finished .
1 *URS: you finished ?
## 2 *URS: all set ?
## 3 *CHI: mine .
## 4 *CHI: my paper .
## 5 *CHI: mine .
6 *MOT: what are you doing to the paper ?
## 7 *CHI: where my pen # Urs(u)la ?
8 *URS: who's that # Adam ?
## 9 *CHI: Cromer .
10 *CHI: let me slip down .
11 *CHI: here we are .
12 *MOT: here we are .
```

Roleset Information
ID : to.Destination
Name : Destination
Arg–P : governor
Arg–P : object

Figure 2: Example from the output of the NLP pipeline parse and semantic role labeler. The label 'auto', is in red, indicating that the labels are the output of the NLP pipeline.

instructed to consult these guidelines for all decisions. Additional decisions that came up are listed in the specifications.[11] Annotators were instructed to make a single-pass of the data. If the output parse interfered with labeling semantic roles, annotators were instructed to change the parse according to the Penn Treebank Guidelines.

In order to check verb senses and semantic role labels, an abbreviated version of Propbank (Kingsbury and Palmer, 2002; Gildea and Palmer, 2002; Palmer et al., 2005) labels appeared in the Jubilee window, and annotators were able to use a link to the Propbank entry in the Unified Verb Index (Bonial et al., 2014; Bonial et al., forthcoming), if additional information was needed.

If a verb sense did not appear to be present in Propbank, annotators consulted a list of previous decisions and, if it was present there, used that decision on semantic role labels. If the verb sense was neither in Propbank nor the list of previous decisions, annotators made note of the sense. However, annotators were instructed to try to use the previous senses as much as possible.[12] The preposition semantic roles, based on those in Srikumar and Roth (2011), had brief example descriptions in the Frameset View window.

4.1 Special Issues in Annotation

The most common issues that came up in annotation involved labeling partial expressions and ambiguous arguments. Because the child speech was labeled, it was often unclear whether expressions were filling semantic roles of a predicate or not and, if they were, it was not clear which role they were filling.

Annotators were instructed to only give a semantic role if it was clear what it was. They were instructed to use only the visible context window at the bottom of the Jubilee tool, which showed about twelve sentences with the target sentence in the middle of the context window, highlighted in blue (see Figure 2) to help determine verb or preposition senses that were uncertain.

[11] The full specifications are available at https://github.com/CogComp/child-discourse-SRL/tree/master/specs.

[12] Additional decisions that came up are listed in the specifications available at https://github.com/CogComp/child-discourse-SRL/tree/master/specs.

5 Inter-Annotator Agreement

The following four files were held-out data: Adam11, Adam15, Adam17, and Adam19. The data in these files were not discussed. Annotators were instructed to annotate according to the same methods that they used in other non-held-out data. There were 27,380 total sentences for annotation, and the held-out data total 5,804, so nearly one-fifth of the data was held-out for inter-annotator agreement. The files were annotated sequentially according to their numbers. After the IAA scores were measured on Adam 11, the results were discussed and used for calibration on subsequent held-out corpora.[13] The data was annotated by two annotators. Annotator X was an undergraduate student in linguistics and a bi-lingual speaker of English and Spanish. Annotator Y was a graduate student in linguistics and a native speaker of English.

File	total annotator X	total annotator Y	original sentences
Adam 11	1159	1133	1403
Adam 15	909	884	1187
Adam 17	1402	1379	1453
Adam 19	1704	1695	1761

Table 1: Table of total annotations of held-out data for each annotator (annotator *'X'* and annotator *'Y'*). The total annotations for each annotators are the annotations in the final annotated file. The original sentences total is the number of sentences that appeared in the Jubilee window to be annotated. Some were removed due to errors in automatic detection of prepositions and verbs or because they were instances of the child repeating one verb many times.

Table 1 shows the total annotations for each annotator and the starting number of annotations in the task. In the course of annotation, the number of annotations in the automatically annotated corpus and previous BabySRL corpus changes. Annotations are added when there is a preposition or verb in the sentence that the automated annotator missed entirely. Annotations are removed for reasons given in the specifications. For example, if a child utterance repeats the verb *'jump'* seven times, the automatic annotation will give seven entries. We decided to only use one verb in a set of repetitions. Annotations are also removed due to the fact that automatic tagging of modal auxiliaries and auxiliaries *'do'* and *'be'* occurred, but the specifications said not to annotate them.

In the subsequent sub-sections, the IAA is broken down into component parts.

Table 2 provides the inter-annotator agreement measures (raw score and Cohen's *kappa*). Each of the four held-out files are presented, followed by an average of the scores across all of the held-out data. For each file, the data is separated by verb SRL and preposition SRL, followed by the agreement on both, labeled *'predicate'*. For each of the three categories, the scores are further separated into the scores on child utterances and the scores on adult utterances, with the scores on child and adult utterances combined provided as well.

Column labels of Table 2 are the various annotations that were measured. The label *'sense'* refers to the Propbank sense, in the case of verbs, and the Srikumar and Roth (2011) sense in the case of prepositions. The label *'# args'* represents the measure of annotator agreement on how many semantic role arguments were present in the sentence. The *'label id'* measures how often annotators provided the same label name to an argument, and the *'span'* column checks annotator agreement on the span of the expression to which a label was assigned.

The syntactic structures are not explicitly compared, however, if the trees are significantly different between annotators, then it is reflected in the span of the arguments, in some cases.

[13] Non-held-out data was done up to Adam 7 at this point, and annotators were calibrating non-held-out data during the process. Adam 1-5 were later re-annotated due to a post-processing error. The work-flow time-line can be found at `https://docs.google.com/spreadsheets/d/19wJtQjt6D4CtoQH4drrpdsw4jiSkzj24U9L4W7Ob4vc/edit?usp=sharing`

Adam 11		sense	sense κ	# args	# args κ	label id	label id κ	span	span κ
Verb	Adult	*	*	91.4	86.7	93.5	91.8	96.8	96.7
	Child	90.4	59.2	86.2	76	88.1	83.1	92.4	92.3
	Total	90.4	59.2	88.8	81.35	90.8	87.45	94.6	94.5
Preposition	Adult	72.8	69.1	94.3	75.3	89.3	84.7	88.3	88.1
	Child	78.8	74.7	84.4	69.7	91.1	87.1	90.2	89.9
	Total	75.8	71.9	89.35	72.5	90.2	85.9	89.25	89
Predicates	Total	92.5	71.8	89	76.9	90.5	86.7	91.9	91.8

Adam 15		sense	sense κ	# args	# args κ	label id	label id κ	span	span κ
Verb	Adult	*	*	92.7	88.5	95	93.6	96.5	96.4
	Child	92.2	73.9	93.6	89.5	95.6	93.8	95.7	95.6
	Total	92.2	73.9	93.2	89	95.3	93.7	96.1	96
Preposition	Adult	73.4	70	96.8	80.6	96.3	94.6	96.3	96.2
	Child	72.4	67	86.2	62.4	92.6	89.3	90.8	90.5
	Total	72.9	68.5	91.5	71.5	94.5	92	93.6	93.4
Predicates	Total	82.6	71.2	92.3	80.3	94.9	92.8	94.8	94.7

Adam 17		sense	sense κ	# args	# args κ	label id	label id κ	span	span κ
Verb	Adult	92.6	71.8	91.4	86.2	93.7	92.1	95.9	95.8
	Child	92.9	63.9	88.4	81.5	93.2	90.5	93.5	93.4
	Total	92.8	67.9	90	83.9	93.5	91.3	94.7	94.6
Preposition	Adult	80.9	78.4	92.9	71	94.4	91.8	93.9	93.7
	Child	85.4	82.6	82.2	54.5	92.8	89.6	90.8	90.6
	Total	83.2	80.5	87.6	62.8	93.6	90.7	92.4	92.2
Predicates	Total	88	74.2	88.7	73.3	93.5	91	93.5	93.4

Adam 19		sense	sense κ	# args	# args κ	label id	label id κ	span	span κ
Verb	Adult	92.6	77.6	90.2	85.2	94.8	93.5	95.7	95.6
	Child	94.2	78.3	92.8	87.4	95.2	93.4	96.5	96.5
	Total	93.4	78	91.5	86.3	95	93.5	96.1	96.1
Preposition	Adult	77.6	74.2	95.6	75.1	96.1	94.3	95.3	95.2
	Child	73.6	70.6	92.4	69.4	95.8	93.8	95.5	95.4
	Total	75.6	72.4	94	72.3	96	94.1	95.4	95.3
Predicates	Total	84.5	75.2	92.8	79.3	95.5	93.8	95.8	95.7

All Files		sense	sense κ	# args	# args κ	label id	label id κ	span	span κ
Verb	Adult	92.6	74.7	91.4	86.7	94.3	92.8	94.5	94.5
	Child	92.4	68.8	90.3	83.6	93	90.2	96.2	96.1
	Total	92.5	71.8	90.8	85.1	93.6	91.5	95.4	95.3
Preposition	Adult	76.2	72.9	94.9	75.5	94	91.4	93.5	93.3
	Child	77.6	73.7	86.3	64	93.1	90	91.9	91.6
	Total	76.9	73.3	90.6	69.8	93.6	90.7	92.6	92.5
Predicates	Total	84.7	72.6	90.7	77.4	93.6	91.1	94	93.9

Table 2: For each of the held-out data sets, the IAA is presented for the verb SRL and the preposition SRL separately and for all SRL predicates (verbs and preposition SRL combined). The IAA is presented for adult language data, for child language data, and for both combined. The chart gives the raw IAA score followed by Cohen's κ for each of the following categories: *sense* gives the agreement on the Propbank sense, in the case of verbs, and the Srikumar and Roth (2011) sense, in the case of prepositions. *# arg* gives the IAA on the number of arguments annotators assigned to each predicate. The *label id* measures the agreement on the name of the label given, provided both annotators gave an expression a semantic role label. The *span* gives agreement on the span of a semantic role argument. The asterisks indicate sense data that we did not count due to a misunderstanding in the specifications, which was corrected.

5.1 Analysis

The IAA reports show a high overall agreement on the held-out dataset.[14]

The verb sense agreement κ scores for adults are much higher in the later half of the annotations. The scores for Adam 11 and 15 (not included) were 24.4 and 25.5, going up in Adam 17 and 19 to 71.8 and 77.6. The initial low scores were due to the two annotators following different specifications. The initial specifications said not to add a Propbank sense to expressions imported from the previous corpus. One annotator was adding senses anyway, and this fact was not discovered until the annotators discussed the IAA results of Adam 11. At that time, the annotator who was not adding senses had already completed annotating Adam 15. At meetings, it was decided that the specifications should be changed to say that we would add senses to verbs imported from the previous corpus. We corrected this in annotations that were finished.

The sense agreement scores for child speech and for prepositions are entirely new to this corpus, as are some of the verbs in adult speech, including the copula verb. Overall, the scores reflect high annotator agreement, given the new data. The overall sense agreement κ score is 72.6.

The argument number is simply a calculation of how many arguments the annotators assigned roles to, regardless of whether or not they assigned those arguments the same label or the same span. The effects of the κ score are more drastic on prepositions because the prepositions can only have at most two arguments, whereas verbs, in contrast, can have five. The κ scores for prepositions also tend to be the lowest due to the frequent absence of arguments and ambiguity of expressions. For example, in the child language phrase *'dog on the floor'* the preposition *'on'* governs *'the floor'*, but *'dog'* may or may not be a governor, depending on how the annotator interprets the child's phrase, given the context window. The noun *'dog'* can be considered to be a governor, similar to the sentence *'I meant the dog on the floor'*, or it can be read as being like the sentence *'The dog is on the floor'*, in which case, the governor is *'is'*, and that verb is missing in the child utterance. Because of such problems, a lower agreement occurs on prepositions overall 69.8κ, as opposed to verbs at 85.1κ.

The label identification matching has very high agreement scores, staying over 90κ throughout most of the annotation. The measure checks how often annotators assigned the same label, given that they assigned a label to an argument. Because this project was concerned with child language acquisition and identification of agents and patients in child speech, this was a welcome result. The agent and patient roles, defined in Propbank guidelines, based on a prototype view of semantic roles (Dowty, 1991), are the most common semantic roles associated with verbs in Propbank. Seeing that annotators consistently agreed on how to assign them in a corpus with high ambiguity attests to the specification development in this area and the feasibility of labeling semantic roles in child speech.

The argument span labels also have scores over 90κ in most of the data. Annotators were instructed to alter the parse according to Penn Treebank guidelines only when the parse was incorrect in a way that altered the span of a semantic role argument. Indirectly, the high agreement on argument span indicates a high agreement on decisions regarding changes to the automatic parser output.

Among the areas that had lower agreement, the agreement scores reflect reasonable variation in the interpretation of data that is frequently ambiguous. For example, when the child says *'go mommy'*, it can be read as an imperative telling his mother to leave or as a statement of the child's intention to go to his mother (e.g., 'I am going to mommy.').

In comparison to previous work, according to `http://cogcomp.org/Data/BabySRL.html`, 15 of the 133 files were held-out for measuring IAA. Across all of the files, annotators agreed on an average of 96.57% of the annotated arguments for span and label. Our span and label average is 93.8%. Considering the increased difficulty of SRLs in child speech, the results are comparable.

[14] The code used to calculate IAA is available at `https://gitlab-beta.engr.illinois.edu/babysrl-group/jubilee/blob/master/src/jubilee/agreement/AgreementCalculator.java`.

6 Availability and Licensing of the Corpus

All uses of the CHILDES corpus have general licensing requirements.[15] For this derived corpus, also cite this document.

The xml version is available for download at `http://cogcomp.org/page/resource_view/115`. The CHAT version is available through TalkBank at `https://childes.talkbank.org/derived/`.

7 Conclusion

In this work, we have described a new resource for NLP studies. It applies a commonly used gold standard for shallow semantic labeling, Propbank, as well as a preposition SRL to adult-child interactions. We explained the methods and tools needed for completing the annotation project, and provided support for the quality of the annotation through inter-annotator agreement measures on held-out data.

Acknowledgements

The BabySRL project is funded by NIH grant R01-HD054448 and NSF grant BCS-060257. We would like to thank the anonymous mentor who gave us excellent feedback on our earlier draft, and the three anonymous reviewers who gave helpful comments.

References

Collin F Baker, Charless J Fillmore, and John B Low. 1998. The Berkeley FrameNet project. *COLING-ACL '98: Proceedings of the Conference*, Montreal, Canada.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II style Penn Treebank Project. Technical report, Linguistic Data Consortium.

Claire Bonial, 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.

Claire Bonial, Kathryn Conger, Jena Hwang, Aous Mansouri, Yahya Aseri, Julia Bonn, Tim O'Gorman, and Martha Palmer. forthcoming. Current directions in English and Arabic Propbank. In *The Handbook of Linguistic Annotations*.

R. Brown. 1973. *A First Language: The Early Stages*. Cambridge: Harvard University Press.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. 2010. Starting from scratch in semantic role labeling. *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, Uppsala, Sweden.

Michael Connor. 2011. *Minimal Supervision for Language Learning: Bootstrapping Global Patterns from Local Knowledge*. PhD thesis, University of Illinois at Urbana-Champaign.

Michael Connor, Cynthia Fisher, and Dan Roth. 2012. Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision. *Cognitive Aspects of Computational Language Acquisition*.

Jino D. Choi, Claire Bonial, and Martha Palmer. 2010. Propbank Instance Annotation Guidelines Using a Dedicated Editor, Jubilee *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

David Dowty. 1991. Thematic Proto-roles and Argument Selection. *Language* 67, 547-619.

Anne Fernald, Michael Frank, Noah Goodman, and Joshua Tenenbaum. 2009. Continuity of discourse provides information for word learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Cynthia Fisher, Yael Gertner, Rose M Scott, and Sylvia Yuan. 2010. Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):143–149, March.

Yael Gertner and Cynthia Fisher. 2012. Predicted errors in children's early sentence comprehension. *Cognition. 2012 July; 124(1):85-94.*

[15] The general usage guidelines are found at `https://talkbank.org/share/rules.html`.

Yael Gertner, Cynthia Fisher, J. Eisengart. 2006. Learning words and rules. *Psychological Science*, 17.

Dan Gildea and Martha Palmer. 2002. The necesity of parsing for predicate argument recognition. In *Proceedings of the ACL 2002*, Philadelphia, Pennsylvania.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain.

Jill Lany and Jenny Saffran. 2010 From statistics to meaning: Infant's acquisition of lexical categories. *Psychological Science* Feb 21(2).

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk.* Lawrence Erlbaum Associates, Inc., 3rd edition.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating Noun Argument Structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004).*

Marcus P. Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2).

L. Pearl and J. Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20, 23-68.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1).

K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of child language* 37(3), 705-729.

K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition* 25-32. Association for Computational Linguistics.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision).

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Kathryn Conger, Tim O'Gorman, and Martha Palmer. 2016. A corpus of preposition supersenses in English web reviews In *Computing Research Repository (CoRR)*

Rushen Shi and Andreane Melancon. 2010. Syntactic categorization in French-learning infants. *Infancy*, 15(5).

Vivek Srikumar and Dan Roth. 2011. A joint model for extended semantic role labeling. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 129-139.

Vivek Srikumar 2013. The semantics of role labeling. Unpublished dissertation. University of Illinois at Urbana-Champaign.

Colin Warner, Arrick Lanfranchi, Tim O'Gorman, Amanda Howard, Kevin Gould, and Michael Regan. 2012. Bracketing biomedical text: An addendum to Penn Treebank II Guidelines. Technical report, Institute of Cognitive Science, University of Colorado at Boulder.