Bootstrapping into Filler-Gap: An Acquisition Story

Marten van Schijndel **Micha Elsner**
The Ohio State University
{vanschm,melsner}@ling.ohio-state.edu

Abstract

Analyses of filler-gap dependencies usually involve complex syntactic rules or heuristics; however recent results suggest that filler-gap comprehension begins earlier than seemingly simpler constructions such as ditransitives or passives. Therefore, this work models filler-gap acquisition as a byproduct of learning word orderings (e.g. SVO vs OSV), which must be done at a very young age anyway in order to extract meaning from language. Specifically, this model, trained on part-of-speech tags, represents the preferred locations of semantic roles relative to a verb as Gaussian mixtures over real numbers.

This approach learns role assignment in filler-gap constructions in a manner consistent with current developmental findings and is extremely robust to initialization variance. Additionally, this model is shown to be able to account for a characteristic error made by learners during this period (*A and B gorped* interpreted as *A gorped B*).

1 Introduction

The phenomenon of filler-gap, where the argument of a predicate appears outside its canonical position in the phrase structure (e.g. *[the apple]$_i$ that the boy ate t_i* or *[what]$_i$ did the boy eat t_i*), has long been an object of study for syntacticians (Ross, 1967) due to its apparent processing complexity. Such complexity is due, in part, to the arbitrary length of the dependency between a filler and its gap (e.g. *[the apple]$_i$ that Mary said the boy ate t_i*).

Recent studies indicate that comprehension of filler-gap constructions begins around 15 months (Seidl et al., 2003; Gagliardi et al., 2014). This finding raises the question of how such a complex phenomenon could be acquired so early since children at that age do not yet have a very advanced grasp of language (e.g. ditransitives do not seem to be generalized until at least 31 months; Goldberg et al. 2004, Bello 2012). This work shows that filler-gap comprehension in English may be acquired through learning word orderings rather than relying on hierarchical syntactic knowledge.

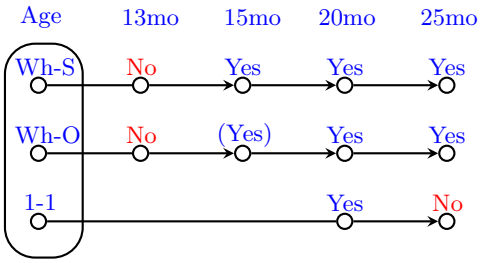

Figure 1: The developmental timeline of subject (Wh-S) and object (Wh-O) *wh*-clause extraction comprehension suggested by experimental results (Seidl et al., 2003; Gagliardi et al., 2014). Parentheses indicate weak comprehension. The final row shows the timeline of 1-1 role bias errors (Naigles, 1990; Gertner and Fisher, 2012). Missing nodes denote a lack of studies.

This work describes a cognitive model of the developmental timecourse of filler-gap comprehension with the goal of setting a lower bound on the modeling assumptions necessary for an ideal learner to display filler-gap comprehension. In particular, the model described in this paper takes chunked child-directed speech as input and learns orderings over semantic roles. These orderings then permit the model to successfully resolve filler-gap dependencies.[1] Further, the model presented here is also shown to initially reflect an idiosyncratic role assignment error observed in development (e.g. *A and B kradded* interpreted as *A kradded B*; Gertner and Fisher, 2012), though after training, the model is able to avoid the error. As such, this work may be said to model a learner from 15 months to between 25 and 30 months.

[1]This model does not explicitly learn gap positions, but rather assigns thematic roles to arguments based on where those arguments are expected to manifest. This approach to filler-gap comprehension is supported by findings that show people do not actually link fillers to gap positions but instead link the filler to a verb with missing arguments (Pickering and Barry, 1991)

2 Background

The developmental timeline during which children acquire the ability to process filler-gap constructions is not well-understood. Language comprehension precedes production, and the developmental literature on the acquisition of filler-gap constructions is sparsely populated due to difficulties in designing experiments to test filler-gap comprehension in preverbal infants. Older studies typically looked at verbal children and the mistakes they make to gain insight into the acquisition process (de Villiers and Roeper, 1995).

Recent studies, however, indicate that filler-gap comprehension likely begins earlier than production (Seidl et al., 2003; Gagliardi and Lidz, 2010; Gagliardi et al., 2014). Therefore, studies of verbal children are probably actually testing the acquisition of production mechanisms (planning, motor skills, greater facility with lexical access, etc) rather than the acquisition of filler-gap. Note that these may be related since filler-gap could introduce greater processing load which could overwhelm the child's fragile production capacity (Phillips, 2010).

Seidl et al. (2003) showed that children are able to process *wh*-extractions from subject position (e.g. *[who]$_i$ t_i ate pie*) as young as 15 months while similar extractions from object position (e.g. *[what]$_i$ did the boy eat t_i*) remain unparseable until around 20 months of age.[2] This line of investigation has been reopened and expanded by Gagliardi et al. (2014) whose results suggest that the experimental methodology employed by Seidl et al. (2003) was flawed in that it presumed infants have ideal performance mechanisms. By providing more trials of each condition and controlling for the pragmatic felicity of test statements, Gagliardi et al. (2014) provide evidence that 15-month old infants can process *wh*-extractions from both subject and object positions. Object extractions are more difficult to comprehend than subject extractions, however, perhaps due to additional processing load in object extractions (Gibson, 1998; Phillips, 2010). Similarly, Gagliardi and Lidz (2010) show that relativized extractions with a *wh*-relativizer (e.g. *find [the boy]$_i$ who t_i ate the apple*) are easier to comprehend than relativized extractions with *that* as the relativizer (e.g. *find [the boy]$_i$ that t_i ate the apple*).

Yuan et al. (2012) demonstrate that 19-month olds use their knowledge of nouns to learn both verbs and their associated argument structure. In their study, infants were shown video of a person talking on a phone using a nonce verb with either one or two nouns (e.g. *Mary kradded Susan*). Under the assumption that infants look longer at things that correspond to their understanding of a prompt, the infants were then shown two images that potentially depicted the described action – one picture where two actors acted independently (reflecting an intransitive proposition) and one picture where one actor acted on the other (reflecting a transitive proposition).[3] Even though the infants had no extralinguistic knowledge about the verb, they consistently treated the verb as transitive if two nouns were present and intransitive if only one noun was present.

Similarly, Gertner and Fisher (2012) show that intransitive phrases with conjoined subjects (e.g. *John and Mary gorped*) are given a transitive interpretation (i.e. *John gorped Mary*) at 21 months (henceforth termed '1-1 role bias'), though this effect is no longer present at 25 months (Naigles, 1990). This finding suggests both that learners will ignore canonical structure in favor of using all possible arguments and that children have a bias to assign a unique semantic role to each argument. It is important to note, however, that cross-linguistically children do not seem to generalize beyond two arguments until after at least 31 months of age (Goldberg et al., 2004; Bello, 2012), so a predicate occurring with three nouns would still likely be interpreted as merely transitive rather than ditransitive.

Computational modeling provides a way to test the computational level of processing (Marr, 1982). That is, given the input (child-directed speech, adult-directed speech, and environmental experiences), it is possible to probe the computational processes that result in the observed output. However, previous computational models of grammar induction (Klein and Manning, 2004), including infant grammar induction (Kwiatkowski et al., 2012), have not addressed filler-gap comprehension.[4]

The closest work to that presented here is the work on BabySRL (Connor et al., 2008; Connor et al., 2009; Connor et al., 2010). BabySRL is a computational model of semantic role acquistion using a similar set of assumptions to the current work. BabySRL learns weights over ordering constraints (e.g. preverbal, second noun, etc.) to acquire semantic role labelling while still exhibiting 1-1 role bias. However, no analysis has evaluated the abil-

[2] Since the *wh*-phrase is in the same (or a very similar) position as the original subject when the *wh*-phrase takes subject position, it is not clear that these constructions are true extractions (Culicover, 2013), however, this paper will continue to refer to them as such for ease of exposition.

[3] There were two actors in each image to avoid biasing the infants to look at the image with more actors.

[4] As one reviewer notes, Joshi et al. (1990) and subsequent work show that filler-gap phenomena can be formally captured by mildly context-sensitive grammar formalisms; these have the virtue of scaling up to adult grammar, but due to their complexity, do not seem to have been described as models of early acquisition.

Susan	said	**John**	gave	**girl**	**book**
-3	-2	-1	0	1	2

Table 1: An example of a chunked sentence (*Susan said John gave the girl a red book*) with the sentence positions labelled. Nominal heads of noun chunks are in bold.

ity of BabySRL to acquire filler-gap constructions. Further comparison to BabySRL may be found in Section 6.

3 Assumptions

The present work restricts itself to acquiring filler-gap comprehension in English. The model presented here learns a single, non-recursive ordering for the semantic roles in each sentence relative to the verb since several studies have suggested that early child grammars may consist of simple linear grammars that are dictated by semantic roles (Diessel and Tomasello, 2001; Jackendoff and Wittenberg, in press). This work assumes learners can already identify nouns and verbs, which is supported by Shi et al. (1999) who show that children at an extremely young age can distinguish between content and function words and by Waxman and Booth (2001) who show that children can distinguish between different types of content words. Further, since Waxman and Booth (2001) demonstrate that, by 14 months, children are able to distinguish nouns from modifiers, this work assumes learners can already chunk nouns and access the nominal head. To handle recursion, this work assumes that children treat the final verb in each sentence as the main verb (implicitly assuming sentence segmentation), which ideally assigns roles to each of the nouns in the sentence.

Due to the findings of Yuan et al. (2012), this work adopts a 'syntactic bootstrapping' theory of acquisition (Gleitman, 1990), where structural properties (e.g. number of nouns) inform the learner about semantic properties of a predicate (e.g. how many semantic roles it confers). Since infants infer the number of semantic roles, this work further assumes they already have expectations about where these roles tend to be realized in sentences, if they appear. These positions may correspond to different semantic roles for different predicates (e.g. the subject of *run* and of *melt*); however, the role for predicates with a single argument is usually assigned to the noun that precedes the verb while a second argument is usually assigned after the verb. The semantic properties of these roles may be learned lexically for each predicate, but that is beyond the scope of this work. Therefore, this work uses syntactic and semantic roles interchangeably (e.g. *subject* and *agent*).

	μ	σ	π
G_{SC}	-1	0.5	.999
G_{SN}	-1	3	.001
G_{OC}	1	0.5	.999
G_{ON}	1	3	.001
Φ	.00001		

Table 2: Initial values for the mean (μ), standard deviation (σ), and prior (π) of each Gaussian as well as the skip penalty (Φ) used in this paper.

Finally, following the finding by Gertner and Fisher (2012) that children interpret intransitives with conjoined subjects as transitives, this work assumes that semantic roles have a one-to-one correspondence with nouns in a sentence (similarly used as a soft constraint in the semantic role labelling work of Titov and Klementiev, 2012).

4 Model

The model represents the preferred locations of semantic roles relative to the verb as distributions over real numbers. This idea is adapted from Boersma (1997) who uses it to learn constraint rankings in optimality theory.

In this work, the final (main) verb is placed at position 0; words (and chunks) before the verb are given progressively more negative positions, and words after the verb are given progressively more positive positions (see Table 1). Learner expectations of where an argument will appear relative to the verb are modelled as two-component Gaussian mixtures: one mixture of Gaussians ($G_{S\cdot}$) corresponds to the subject argument, another ($G_{O\cdot}$) corresponds to the object argument. There is no mixture for a third argument since children do not generalize beyond two arguments until later in development (Goldberg et al., 2004; Bello, 2012).

One component of each mixture learns to represent the canonical position for the argument ($G_{\cdot C}$) while the other ($G_{\cdot N}$) represents some alternate, non-canonical position such as the filler position in filler-gap constructions. To reflect the fact that learners have had 15 months of exposure to their language before acquiring filler-gap, the mixture is initialized so that there is a stronger probability associated with the canonical Gaussian than with the non-canonical Gaussian of each mixture.[5] Finally, the one-to-one role bias is explicitly encoded such that the model cannot use a label that has already been used elsewhere in the sentence.

[5] Akhtar (1999) finds that learners may not have strong expectations of canonical argument positions until four years of age, but the results of the current study are extremely robust to changes in initialization, as discussed in Section 7 of this paper, so this assumption is mostly adopted for ease of exposition.

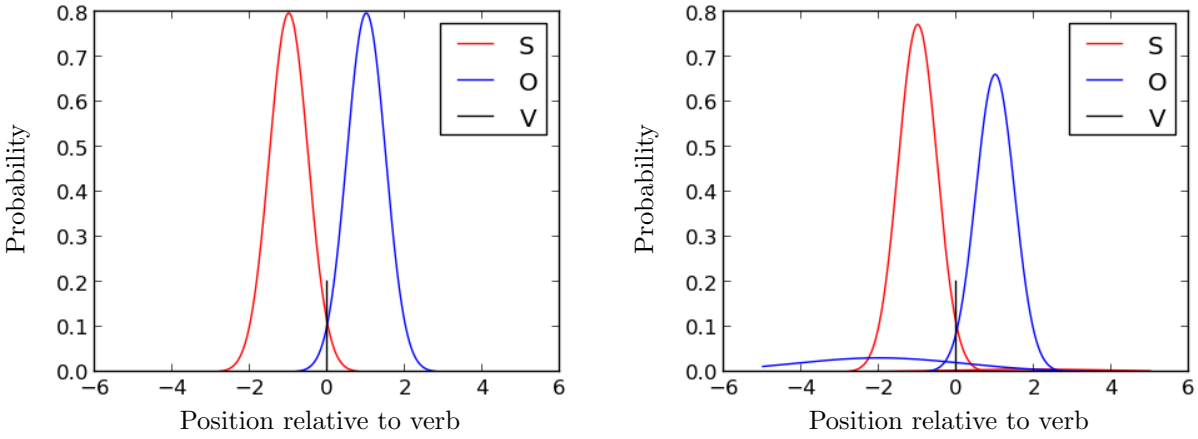

Figure 2: Visual representations of (Left) the initial model's expectations of where arguments will appear, given the initial parameters in Table 2 and (Right) the converged model's expectations of where arguments will appear.

Thus, the initial model conditions (see Figure 2) are most likely to realize an SVO ordering, although it is possible to obtain SOV (by sampling a negative number from the blue curve) or even OSV (by also sampling the red curve very close to 0). The model is most likely to hypothesize a preverbal object when it has already assigned the subject role to something and, in addition, there is no postverbal noun competing for the object label. In other words, the model infers that an object extraction may have occurred if there is a 'missing' postverbal argument.

Finally, the probability of a given sequence is the product of the label probabilities for the component argument positions (e.g. G_{SC} generating an argument at position -2, etc). Since many sentences have more than two nouns, the model is allowed to skip nouns by multiplying a penalty term (Φ) into the product for each skipped noun; the cost is set at 0.00001 for this study, though see Section 7 for a discussion of the constraints on this parameter. See Table 2 for initialization parameters and Figure 2 for a visual representation of the initial expectations of the model.

This work uses a model with 2-component mixtures for both subjects and objects (termed the *symmetric model*). This formulation achieves the best fit to the training data according to the Bayesian Information Criterion (BIC).[6] However, follow-up experiments find that the non-canonical subject Gaussian only improves the likelihood of the data by erroneously modeling postverbal nouns in imperative statements. The lack of a canonical subject in English imperatives allows the model to improve the likelihood of the data by using the non-canonical subject Gaussian to capture fictitious postverbal arguments. When imperatives are filtered out of the training corpus, the symmetric model obtains a worse BIC fit than a model that lacks the non-canonical subject Gaussian. Therefore, if one makes the assumption that imperatives are prosodically-marked for learners (e.g. the learner is the implicit subject), the best model is one that lacks a non-canonical subject.[7] The remainder of this paper assumes a symmetric model to demonstrate what happens if such an assumption is not made; for the evaluations described in this paper, the results are similar in either case.

This model differs from other non-recursive computational models of grammar induction (e.g. Goldwater and Griffiths, 2007) since it is not based on Hidden Markov Models. Instead, it determines the best ordering for the sentence as a whole. This approach bears some similarity to a Generalized Mallows model (Chen et al., 2009), but the current formulation was chosen due to being independently posited as cognitively plausible (Boersma, 1997).

Figure 2 (Right) shows the converged, final state of the model. The model expects the first argument (usually agent) to be assigned preverbally and expects the second (say, patient) to be assigned postverbally; however, there is now a larger chance that the second argument will appear preverbally.

5 Evaluation

The model in this work is trained using transcribed child-directed speech (CDS) from the BabySRL portions (Connor et al., 2008) of CHILDES (MacWhinney, 2000). Chunking is performed us-

[6]The BIC rewards improved log-likelihood but penalizes increased model complexity.

[7]This finding suggests that a Dirichlet Process or other means of dynamically determining the number of components in each mixture would converge to a model that lacks non-canonical subjects if imperative filtering were employed.

	Eve (n = 4820)			Adam (n = 4461)		
	P	R	F	P	R	F
Initial	.54	.64	.59	.53	.60	.56
Trained	.52	.69	.59*	.51	.65	.57*
Initial$_c$	.56	.66	.60	.55	.62	.58
Trained$_c$	.54	.71	.61*	.53	.67	.59*

Table 3: Overall accuracy on the Eve and Adam sections of the BabySRL corpus. Bottom rows reflect accuracy when non-agent roles are collapsed into a single role. Note that improvements are numerically slight since filler-gap is relatively rare (Schuler, 2011). $^*p << .01$

Subject Extraction filter: S x V ...
Object Extraction filter: O ... V ...

	Eve (n = 1345)			Adam (n = 1287)		
	P	R	F	P	R	F
Initial$_c$	.53	.57	.55	.53	.52	.52
Trained$_c$	.55	.67	.61*	.54	.63	.58*

Table 4: (Above) Filters to extract filler-gap constructions: A) the subject and verb are not adjacent, B) the object precedes the verb. (Below) Filler-gap accuracy on the Eve and Adam sections of the BabySRL corpus when non-agent roles are collapsed into a single role. $^*p << .01$

ing a basic noun-chunker from NLTK (Bird et al., 2009). Based on an initial analysis of chunker performance, *yes* is hand-corrected to not be a noun. Poor chunker perfomance is likely due to a mismatch in chunker training and testing domains (Wall Street Journal text vs transcribed speech), but chunking noise may be a good estimation of learner uncertainty, so the remaining text is left uncorrected. All noun phrase chunks are then replaced with their final noun (presumed the head) to approximate the ability of children to distinguish nouns from modifiers (Waxman and Booth, 2001). Finally, for each sentence, the model assigns sentence positions to each word with the final verb at zero.

Viterbi Expectation-Maximization is performed over each sentence in the corpus to infer the parameters of the model. During the Expectation step, the model uses the current Gaussian parameters to label the nouns in each sentence with argument roles. Since the model is not lexicalized, these roles correspond to the semantic roles most commonly associated with subject and object. The model then chooses the best label sequence for each sentence.

These newly labelled sentences are used during the Maximization step to determine the Gaussian parameters that maximize the likelihood of that labelling. The mean of each Gaussian is updated to the mean position of the words it labels. Similarly, the standard deviation of each Gaussian is updated with the standard deviation of the positions it labels. A learning rate of 0.3 is used to prevent large parameter jumps. The prior probability of each Gaussian is updated as the ratio of that Gaussian's labellings to the total number of labellings from that mixture in the corpus:

$$\pi_{\rho\theta} = \frac{|\ G_{\rho\theta}\ |}{|\ G_{\rho\cdot}\ |} \quad (1)$$

where $\rho \in \{S, O\}$ and $\theta \in \{C, N\}$.

Best results seem to be obtained when the skip-penalty is loosened by an order of magnitude during testing. Essentially, this forces the model to tightly adhere to the perceived argument structure during training to learn more rigid parameters, but the model is allowed more leeway to skip arguments it has less confidence in during testing. Convergence (see Figure 2) tends to occur after four iterations but can take up to ten iterations depending on the initial parameters.

Since the model is unsupervised, it is trained on a given corpus (e.g. Eve) before being tested on the role annotations of that same corpus. The Eve corpus was used for development purposes,[8] and the Adam data was used only for testing.

For testing, this study uses the semantic role annotations in the BabySRL corpus. These annotations were obtained by automatically semantic role labelling portions of CHILDES with the system of Punyakanok et al. (2008) before roughly hand-correcting them (Connor et al., 2008). The BabySRL corpus is annotated with 5 different roles, but the model described in this paper only uses 2 roles. Therefore, overall accuracy results (see Table 3) are presented both for the raw BabySRL corpus and for a collapsed BabySRL corpus where all non-agent roles are collapsed into a single role (denoted by a subscript $_c$ in all tables).

Since children do not generalize above two arguments during the modelled age range (Goldberg et al., 2004; Bello, 2012), the collapsed numbers more closely reflect the performance of a learner at this age than the raw numbers. The increase in accuracy obtained from collapsing non-agent arguments indicates that children may initially generalize incorrectly to some verbs and would need to learn lexically-specific role assignments (e.g. double-object constructions of *give*). Since the current work is interested in general filler-gap comprehension at this age, including over unknown verbs, the remaining analyses in this paper con-

[8]This is included for transparency, though the initial parameters have very little bearing on the final results as stated in Section 7, so the danger of overfitting to development data is very slight.

	P	R	F	P	R	F
Eve	Subj (n = 691)			Obj (n = 654)		
Initial_c	.66	.83	.74	.35	.31	.33
Trained_c	.64	.84	$.72^{\dagger}$	.45	.52	$.48^{*}$
Adam	Subj (n = 886)			Obj (n = 1050)		
Initial_c	.69	.81	.74	.33	.27	.30
Trained_c	.66	.81	.73	.44	.48	$.46^{*}$

	P	R	F	P	R	F
Eve	Wh- (n = 689)			That (n = 125)		
Initial_c	.63	.45	.53	.43	.48	.45
Trained_c	.73	.75	$.74^{*}$	.44	.57	$.50^{\dagger}$
Adam	Wh- (n = 748)			That (n = 189)		
Initial_c	.50	.37	.42	.50	.50	.50
Trained_c	.61	.65	$.63^{*}$	.47	.56	$.51^{\dagger}$

Table 5: (Left) Subject-extraction accuracy and object-extraction accuracy and (Right) *Wh*-relative accuracy and *that*-relative accuracy; calculated over the Eve and Adam sections of the BabySRL corpus with non-agent roles collapsed into a single role. $^{\dagger}p = .02$ $^{*}p << .01$

sider performance when non-agent arguments are collapsed.[9]

Next, a filler-gap version of the BabySRL corpus is created using a coarse filtering process: the new corpus is comprised of all sentences where an associated object precedes the final verb and all sentences where the relevant subject is not immediately followed by the final verb (see Table 4). For these filler-gap evaluations, the model is trained on the full version of the corpus in question (e.g. Eve) before being tested on the filler-gap subset of that corpus. The overall results of the filler-gap evaluation (see Table 4) indicate that the model improves significantly at parsing filler-gap constructions after training.

The performance of the model on role-assignment in filler-gap constructions may be analyzed further in terms of how the model performs on subject-extractions compared with object-extractions and in terms of how the model performs on *that*-relatives compared with *wh*-relatives (see Table 5).

The model actually performs worse at subject-extractions after training than before training. This is unsurprising because, prior to training, subjects have little-to-no competition for preverbal role assignments; after training, there is a preverbal extracted object category, which the model can erroneously use. This slight, though significant in Eve, deficit is counter-balanced by a very substantial and significant improvement in object-extraction labelling accuracy.

Similarly, training confers a large and significant improvement for role assignment in *wh*-relative constructions, but it yields less of an improvement for *that*-relative constructions. This difference mimics a finding observed in the developmental literature where children seem slower to acquire comprehension of *that*-relatives than of *wh*-relatives (Gagliardi and Lidz, 2010).

[9]Though performance is slightly worse when arguments are not collapsed, all the same patterns emerge.

6 Comparison to BabySRL

The acquisition of semantic role labelling (SRL) by the BabySRL model (Connor et al., 2008; Connor et al., 2009; Connor et al., 2010) bears many similarities to the current work and is, to our knowledge, the only comparable line of inquiry to the current one. The primary function of BabySRL is to model the acquisition of semantic role labelling while making an idiosyncratic error which infants also make (Gertner and Fisher, 2012), the 1-1 role bias error (*John and Mary gorped* interpreted as *John gorped Mary*). Similar to the model presented in this paper, BabySRL is based on simple ordering features such as argument position relative to the verb and argument position relative to the other arguments.

This section will demonstrate that the model in this paper initially reflects 1-1 role bias comparably to BabySRL, though it progresses beyond this bias after training.[10] Further, the model in this paper is able to reflect the concurrent acquisition of filler-gap whereas BabySRL does not seem well-suited to such a task. Finally, BabySRL performs undesirably in intransitive settings whereas the model in this paper does not.

Connor et al. (2008) demonstrate that a supervised perceptron classifier, based on positional features and trained on the silver role label annotations of the BabySRL corpus, manifests 1-1 role bias errors. Follow-up studies show that supervision may be lessened (Connor et al., 2009) or removed (Connor et al., 2010) and BabySRL will still reflect a substantial 1-1 role bias.

Connor et al. (2008) and Connor et al. (2009) run direct analyses of how frequently their models make 1-1 role bias errors. A comparable evaluation may be run on the current model by generating 1000 sentences with a structure of NNV and reporting how many times the model chooses a subject-first labelling (see Table 6).[11]

[10]All evaluations in this section are preceded by training on the chunked Eve corpus.

[11]While Table 6 analyzes erroneous labellings of NNV structure, the 'Obj' column of Table 5 (Left)

	Error rate
Initial	.36
Trained	.11
Initial (given 2 args)	.66
Trained (given 2 args)	.13
2008 arg-arg position	.65
2008 arg-verb position	0
2009 arg-arg position	.82
2009 arg-verb position	.63

Table 6: 1-1 role bias error in this model compared to the models of Connor et al. (2008) and Connor et al. (2009). That is, how frequently each model labelled an NNV sentence SOV. Since the Connor et al. models are perceptron-based, they require both arguments be labelled. The model presented in this paper does not share this restriction, so the raw error rate for this model is presented in the first two lines; the error rate once this additional restriction is imposed is given in the second two lines.

The results of Connor et al. (2008) and Connor et al. (2009) depend on whether BabySRL uses argument-argument relative position as a feature or argument-verb relative position as a feature (there is no combined model). Further, the model presented here from Connor et al. (2009) has a unique argument constraint, similar to the model in this paper, in order to make comparison as direct as possible.

The 1-1 role bias error rate (before training) of the model presented in this paper is comparable to that of Connor et al. (2008) and Connor et al. (2009), which shows that the current model provides comparable developmental modeling benefits to the BabySRL models. Further, similar to real children (see Figure 1) the model presented in this paper develops beyond this error by the end of its training,[12] whereas the BabySRL models still make this error after training.

Connor et al. (2010) look at how frequently their model correctly labels the agent in transitive and intransitive sentences with unknown verbs (to demonstrate that it exhibits an agent-first bias). This evaluation can be replicated for the current study by generating 1,000 sentences with the transitive form of NVN and a further 1,000 sentences with the intransitive form of NV (see Table 7).

Since Connor et al. (2010) investigate the effects of different initial lexicons, this evaluation compares against the resulting BabySRL from each initializer: they initially seed their part-of-speech tagger with either the 10 or 365 most frequent nouns in the corpus or they dispense with the tagger and use gold part-of-speech tags.

	NVN	NV
Sents in Eve	1173	1513
Sents in Adam	1029	1353
Initial	.67	1
Trained	.65	.96
Weak (10) lexical	.71	.59
Strong (365) lexical	.74	.41
Gold Args	.77	.58

Table 7: Agent-prediction recall accuracy in transitive (NVN) and intransitive (NV) settings of the model presented in this paper (middle) and the combined model of Connor et al. (2010) (bottom), which has features for argument-argument relative position as well as argument-predicate relative position and so is closest to the model presented in this paper.

As with subject extraction, the model in this paper gets less accurate after training because of the newly minted extracted object category that can be mistakenly used in these canonical settings. While the model of Connor et al. (2010) outperforms the model presented here when in a transitive setting, their model does much worse in an intransitive setting. The difference in transitive settings stems from increased lexicalization, as is apparent from their results alone; the model presented here initially performs close to their weakly lexicalized model, though training impedes agent-prediction accuracy due to an increased probability of non-canonical objects.

For the intransitive case, however, whereas the model presented in this paper is generally able to successfully label the lone noun as the subject, the model of Connor et al. (2010) chooses to label lone nouns as objects about 40% of the time. This likely stems from their model's reliance on argument-argument relative position as a feature; when there is no additional argument to use for reference, the model's accuracy decreases. This is borne out by their model (not shown in Table 7) that omits the argument-argument relative position feature and solely relies on verb-argument position, which achieves up to 70% accuracy in intransitive settings. Even in that case, however, BabySRL still chooses to label lone nouns as objects 30% of the time. The fact that intransitive sentences are more common than transitive sentences in both the Eve and Adam sections of the BabySRL corpus suggests that learners should be more likely to assign

shows model accuracy on NNV structures.

[12] It is important to note that the unique argument constraint prevents the current model from actually getting the correct, conjoined-subject parse, but it no longer exhibits agent-first bias, an important step for acquiring passives, which occurs between 3 and 4 years (Thatcher et al., 2008).

correct roles in an intransitive setting, which is not reflected in the BabySRL results.

The overall reason for the different results between the current work and BabySRL is that BabySRL relies on positional features that measure the relative position of two individual elements (e.g. where a given noun is relative to the verb). Since the model in this paper operates over global orderings, it implicitly takes into account the positions of other nouns as it models argument position relative to the verb; object and subject are in competition as labels for preverbal nouns, so a preverbal object is usually only assigned once a subject has already been detected.

Further, while BabySRL consistently reflects 1-1 role bias (corresponding to a pre 25-month old learner), it also learns to productively label five roles, which developmental studies have shown does not take place until at least 31 months (Goldberg et al., 2004; Bello, 2012). Finally, it does not seem likely that BabySRL could be easily extended to capture filler-gap acquisition. The argument-verb position features impede acquisition of filler-gap by classifying preverbal arguments as agents, and the argument-argument position features inhibit accurate labelling in intransitive settings and result in an agent-first bias which would tend to label extracted objects as agents. In fact, these observations suggest that any linear classifier which relies on positioning features will have difficulties modeling filler-gap acquisition.

In sum, the unlexicalized model presented in this paper is able to achieve greater labelling accuracy than the lexicalized BabySRL models in intransitive settings, though this model does perform slightly worse in the less common transitive setting. Further, the unsupervised model in this paper initially reflects developmental 1-1 role bias as well as the supervised BabySRL models, and it is able to progress beyond this bias. Finally, unlike BabySRL, the model presented here provides a cognitive model of the acquisition of filler-gap comprehension, which BabySRL does not seem well-suited to model.

7 Discussion

This paper has presented a simple cognitive model of filler-gap acquisition, which is able to capture several findings from developmental psychology. Training significantly improves role labelling in the case of object-extractions, which improves the overall accuracy of the model. This boost is accompanied by a slight decrease in labelling accuracy in subject-extraction settings. The asymmetric ease of subject versus object comprehension is well-documented in both children and adults (Gibson, 1998), and while training improves the model's ability to process object-extractions, there is still a gap between object-extraction and subject-extraction comprehension even after training.

Further, the model exhibits better comprehension of *wh*-relatives than *that*-relatives similar to children (Gagliardi and Lidz, 2010). This could also be an area where a lexicalized model could do better. As Gagliardi and Lidz (2010) point out, whereas *wh*-relatives such as *who* or *which* always signify a filler-gap construction, *that* can occur for many different reasons (demonstrative, determiner, complementizer, etc) and so is a much weaker filler-gap cue. A lexical model could potentially pick up on clues which could indicate when *that* is a relativizer or simply improve on its comprehension of *wh*-relatives even more.

It is interesting to note that the cuurent model does not make use of *that* as a cue at all and yet is still slower at acquiring *that*-relatives than *wh*-relatives. This fact suggests that the findings of Gagliardi and Lidz (2010) may be partially explained by a frequency effect: perhaps the input to children is simply biased such that *wh*-relatives are much more common than *that*-relatives (as shown in Table 5).

This model also initially reflects the 1-1 role bias observed in children (Gertner and Fisher, 2012) as well as previous models (Connor et al., 2008; Connor et al., 2009; Connor et al., 2010) without sacrificing accuracy in canonical intransitive settings.

Finally, this model is extremely robust to different initializations. The canonical Gaussian expectations can begin far from the verb (± 3) or close to the verb (± 0.1), and the standard deviations of the distributions and the skip-penalty can vary widely; the model always converges to give comparable results to those presented here. The only constraint on the initial parameters is that the probability of the extracted object occurring preverbally must exceed the skip-penalty (i.e. extraction must be possible). In short, this paper describes a simple, robust cognitive model of the development of a learner between 15 months until somewhere between 25- and 30-months old (since 1-1 role bias is no longer present but no more than two arguments are being generalized).

In future, it would be interesting to incorporate lexicalization into the model presented in this paper, as this feature seems likely to bridge the gap between this model and BabySRL in transitive settings. Lexicalization should also help further distinguish modifiers from arguments and improve the overall accuracy of the model.

It would also be interesting to investigate how well this model generalizes to languages besides English. Since the model is able to use the verb position as a semi-permeable boundary between canonical subjects and objects, it may not work as

well in verb-final languages, and thus makes the prediction that filler-gap comprehension may be acquired later in development in such languages due to a greater reliance on hierarchical syntax.

Ordering is one of the definining characteristics of a language that must be acquired by learners (e.g. SVO vs SOV), and this work shows that filler-gap comprehension can be acquired as a by-product of learning orderings rather than having to resort to higher-order syntax. Note that this model cannot capture the constraints on filler-gap usage which require a hierarchical grammar (e.g. subjacency), but such knowledge is really only needed for successful production of filler-gap constructions, which occurs much later (around 5 years; de Villiers and Roeper, 1995). Further, the kind of ordering system proposed in this paper may form an initial basis for learning such grammars (Jackendoff and Wittenberg, in press).

8 Acknowledgements

Thanks to Peter Culicover, William Schuler, Laura Wagner, and the attendees of the OSU 2013 Fall Linguistics Colloquium Fest for feedback on this work. This work was partially funded by an OSU Dept. of Linguistics Targeted Investment for Excellence (TIE) grant for collaborative interdisciplinary projects conducted during the academic year 2012-13.

References

Nameera Akhtar. 1999. Acquiring basic word order: evidence for data-driven learning of syntactic structure. *Journal of Child Language*, 26:339–356.

Sophia Bello. 2012. Identifying indirect objects in French: An elicitation task. In *Proceedings of the 2012 annual conference of the Canadian Linguistic Association.*

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly, Beijing.

Paul Boersma. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 21:43–58.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Content modeling using latent permutations. *Journal of Artificial Intelligence Research*, 36:129–163.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. 2008. Baby srl: Modeling early language acquisition. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning.*

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. 2009. Minimally supervised model of early language acquisition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning.*

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. 2010. Starting from scratch in semantic role labelling. In *Proceedings of ACL 2010.*

Peter Culicover. 2013. *Explaining syntax: representations, structures, and computation.* Oxford University Press.

Jill de Villiers and Thomas Roeper. 1995. Barriers, binding, and acquisition of the dp-np distinction. *Language Acquisition*, 4(1):73–104.

Holger Diessel and Michael Tomasello. 2001. The acquisition of finite complement clauses in english: A corpus-based analysis. *Cognitive Linguistics*, 12:1–45.

Annie Gagliardi and Jeffrey Lidz. 2010. Morphosyntactic cues impact filler-gap dependency resolution in 20- and 30-month-olds. In *Poster session of BUCLD35.*

Annie Gagliardi, Tara M. Mease, and Jeffrey Lidz. 2014. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15- and 20-month-olds. Harvard unpublished manuscript: http://www.people.fas.harvard.edu/∼gagliardi.

Yael Gertner and Cynthia Fisher. 2012. Predicted errors in children's early sentence comprehension. *Cognition*, 124:85–94.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Lila R. Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Adele E. Goldberg, Devin Casenhiser, and Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics*, 14(3):289–316.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.*

Ray Jackendoff and Eva Wittenberg. in press. What you can say without syntax: A hierarchy of grammatical complexity. In Fritz Newmeyer and Lauren Preston, editors, *Measuring Linguistic Complexity.* Oxford University Press.

Aravind K. Joshi, K. Vijay Shanker, and David Weir. 1990. The convergence of mildly context-sensitive grammar formalisms. Technical Report MS-CIS-90-01, Department of Computer and Information Science, University of Pennsylvania.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Tom Kwiatkowski, Sharon Goldwater, Luke S. Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of EACL 2012*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Elrbaum Associates, Mahwah, NJ, third edition.

David Marr. 1982. *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Letitia R. Naigles. 1990. Children use syntax to learn verb meanings. *The Journal Child Language*, 17:357–374.

Colin Phillips. 2010. Some arguments and nonarguments for reductionist accounts of syntactic phenomena. *Language and Cognitive Processes*, 28:156–187.

Martin Pickering and Guy Barry. 1991. Sentence processing without empty categories. *Language and Cognitive Processes*, 6(3):229–259.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

John R. Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

William Schuler. 2011. Effects of filler-gap dependencies on working memory requirements for parsing. In *Proceedings of COGSCI*, pages 501–506, Austin, TX. Cognitive Science Society.

Amanda Seidl, George Hollich, and Peter W. Jusczyk. 2003. Early understanding of subject and object wh-questions. *Infancy*, 4(3):423–436.

Rushen Shi, Janet F. Werker, and James L. Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2):B11–B21.

Katherine Thatcher, Holly Branigan, Janet McLean, and Antonella Sorace. 2008. Children's early acquisition of the passive: Evidence from syntactic priming. In *Proceedings of the Child Language Seminar 2007*, pages 195–205, University of Reading.

Ivan Titov and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*.

Sandra R. Waxman and Amy E. Booth. 2001. Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology*, 43:217–242.

Sylvia Yuan, Cynthia Fisher, and Jesse Snedeker. 2012. Counting the nouns: Simple structural cues to verb meaning. *Child Development*, 83(4):1382–1399.