Modeling Child Divergences from Adult Grammar

Sam Sahakian
University of Wisconsin-Madison
sahakian@cs.wisc.edu

Benjamin Snyder
University of Wisconsin-Madison
bsnyder@cs.wisc.edu

Abstract

During the course of first language acquisition, children produce linguistic forms that do not conform to adult grammar. In this paper, we introduce a data set and approach for systematically modeling this child-adult grammar divergence. Our corpus consists of child sentences with corrected adult forms. We bridge the gap between these forms with a discriminatively reranked noisy channel model that translates child sentences into equivalent adult utterances. Our method outperforms MT and ESL baselines, reducing child error by 20%. Our model allows us to chart specific aspects of grammar development in longitudinal studies of children, and investigate the hypothesis that children share a common developmental path in language acquisition.

1 Introduction

Since the publication of the Brown Study (1973), the existence of standard stages of development has been an underlying assumption in the study of first language learning. As a child moves towards language mastery, their language use grows predictably to include more complex syntactic structures, eventually converging to full adult usage. In the course of this process, children may produce linguistic forms that do not conform to the grammatical standard. From the adult point of view these are language errors, a label which implies a faulty production. Considering the work-in-progress nature of a child language learner, these divergences could also be described as expressions of the structural differences between child and adult grammar. The predictability of these divergences has been observed by psychologists, linguists and parents (Owens, 2008).[1]

Our work leverages the differences between child and adult language to make two contributions towards the study of language acquisition. First, we provide a corpus of errorful child sentences annotated with adult-like rephrasings. This data will allow researchers to test hypotheses and build models relating the development of child language to adult forms. Our second contribution is a probabilistic model trained on our corpus that predicts a grammatical rephrasing given an errorful child sentence.

The generative assumption of our model is that sentences begin in underlying adult forms, and are then stochastically transformed into observed child utterances. Given an observed child utterance s, we calculate the probability of the corrected adult translation t as

$$P(t|s) \propto P(s|t)P(t),$$

where $P(t)$ is an adult language model and $P(s|t)$ is a noise model crafted to capture child grammar errors like omission of certain function words and corruptions of tense or declension. The parameters of this noise model are estimated using our corpus of child and adult-form utterances, using EM to handle unobserved word alignments. We use this generative model to produce n-best lists of candidate corrections which are then reranked using long range sentence features in a discriminative framework (Collins and Roark, 2004).

[1] For the remainder of this paper we use "error" and "divergence" interchangeably.

One could argue that our noisy channel model mirrors the cognitive process of child language production by appealing to the hypothesis that children rapidly learn adult-like grammar but produce errors due to performance factors (Bloom, 1990; Hamburger and Crain, 1984). That being said, our primary goal in this paper is not cognitive plausibility, but rather the creation of a practical tool to aid in the empirical study of language acquisition. By automatically inferring adult-like forms of child sentences, our model can highlight and compare developmental trends of children over time using large quantities of data, while minimizing the need for human annotation.

Besides this, our model's predictive success itself has theoretical implications. By aggregating training and testing data across children, our model instantiates the Brown hypothesis of a shared developmental path. Even when adequate per-child training data exists, using data only from other children leads to no degradation in performance, suggesting that the learned parameters capture general child language phenomena and not just individual habits. Besides aggregating across children, our model coarsely lumps together all stages of development, providing a frozen snapshot of child grammar. This establishes a baseline for more cognitively plausible and temporally dynamic models.

We compare our correction system against two baselines, a phrase-based Machine Translation (MT) system, and a model designed for English Second Language (ESL) error correction. Relative to the best performing baseline, our approach achieves a 30% decrease in word error-rate and a four point increase in BLEU score. We analyze the performance of our system on various child error categories, highlighting our model's strengths (correcting *be* drops and morphological overgeneralizations) as well as its weaknesses (correcting pronoun and auxiliary drops). We also assess the learning rate of our model, showing that very little annotation is needed to achieve high performance. Finally, to showcase a potential application, we use our model to chart one aspect of four children's grammar acquisition over time. While generally vindicating the Brown thesis of a common developmental path, the results point to subtleties in variation across individuals that merit further investigation.

2 Background and Related Work

While child error correction is a novel task, computational methods are frequently used to study first language acquisition. The computational study of speech is facilitated by TalkBank (MacWhinney, 2007), a large database of transcribed dialogues including CHILDES (MacWhinney, 2000), a subsection composed entirely of child conversation data. Computational tools have been developed specifically for the large-scale analysis of CHILDES. These tools enable further computational study such as the automatic calculation of the language development metrics IPSYN (Sagae et al., 2005) and D-Level (Lu, 2009), or the automatic formulation of novel language development metrics themselves (Sahakian and Snyder, 2012).

The availability of child language is also key to the design of computational models of language learning (Alishahi, 2010), which can support the plausibility of proposed human strategies for tasks like semantic role labeling (Connor et al., 2008) or word learning (Regier, 2005). To our knowledge this paper is the first work on error correction in the first language learning domain. Previous work has employed a classifier-based approach to identify speech errors indicative of language disorders in children (Morley and Prud'hommeaux, 2012).

Automatic correction of second language (L2) writing is a common objective in computer assisted language learning (CALL). These tasks generally target high-frequency error categories including article, word-form, and preposition choice. Previous work in CALL error correction includes identifying word choice errors in TOEFL essays based on context (Chodorow and Leacock, 2000), correcting errors with a generative lattice and PCFG reranking (Lee and Seneff, 2006), and identifying a broad range of errors in ESL essays by examining linguistic features of words in sequence (Gamon, 2011). In a 2011 shared ESL correction task (Dale and Kilgarriff, 2011), the best performing system (Rozovskaya et al., 2011) corrected preposition, article, punctuation and spelling errors by building classifiers for each category. This line of work is grounded in the practical application of automatic error correction as a learning tool for ESL students.

Statistical Machine Translation (SMT) has been

applied in diverse contexts including grammar correction as well as paraphrasing (Quirk et al., 2004), question answering (Echihabi and Marcu, 2003) and prediction of twitter responses (Ritter et al., 2011). In the realm of error correction, SMT has been applied to identify and correct spelling errors in internet search queries (Sun et al., 2010). Within CALL, Park and Levy (2011) took an unsupervised SMT approach to ESL error correction using Weighted Finite State Transducers (FSTs). The work described in this paper is inspired by that of Park and Levy, and in Section 6 we detail differences between our approaches. We also include their model as a baseline.

3 Data

To train and evaluate our translation system, we first collected a corpus of 1,000 errorful child-language utterances from the American English portion of the CHILDES database. To encourage diversity in the grammatical divergences captured by our corpus, our data is drawn from a large pool of studies (see bibliography for the full list of citations).

In the annotation process, candidate child sentences were randomly selected from the pool and classified by hand as either grammatically correct, divergent or unclassifiable (when it was not possible to tell what a child is trying to say). We continued this process until 1,000 divergent sentences were found. Along the way we also encountered 5,197 grammatically correct utterances and 909 that were unclassifiable.[2] Because CHILDES includes speech samples from children of diverse age, background and language ability, our corpus does not capture any specific stage of language development. Instead, the corpus represents a general snapshot of a learner who has not yet mastered English as their first language.

To provide the grammatically correct counterpart to child data, our errorful sentences were corrected by workers on Amazon's Mechanical Turk web service. Given a child utterance and its surrounding conversational context, annotators were instructed to translate the child utterance into adult-like English. We limited eligible workers to native English speakers residing in the US. We also required annotators to follow a brief tutorial in which they practice correcting sample utterances according to our guidelines. These guidelines instructed workers to minimally alter sentences to be grammatically consistent with a conversation or written letter, without altering underlying meaning. Annotators were evaluated on a worker-by-worker basis and rejected in the rare case that they ignored our guidelines. Accepted workers were paid 7 cents for correcting each set of 5 sentences. To achieve a consistent judgment, we posted each set of sentences for correction by 7 different annotators.

Error Type	Child Utterance
Insertion	I did locked it.
Inflection	More cookie?
Deletion	That not how.
Lemma Choice	I got grain.
Overgeneralization	I drawed it.

Table 1: Examples of error types captured by our model.

Once multiple reference translations were obtained we selected a single best correction by plurality, arbitrating ties as necessary. There were several cases in which corrections obtained by plurality decision did not perfectly follow instructions. These were manually corrected. Both the raw translations provided by individual annotators as well as the curated final adult forms are provided online as part of our data set.[3] Resulting pairs of errorful child sentences and their adult-like corrections were split into 73% training, 7% development and 20% test data, which we use to build, tune and evaluate our grammar correction system. In the final test phase, development data is included in the training set.

4 Model

According to our generative model, adult-like utterances are formed and then transformed by a noisy channel to become child sentences. The structure of our noise model is tailored to match our observations of common child errors. These include: function word insertions, function word deletions, swaps of function words and, inflectional changes to content words. Examples of each error type are given

[2] These hand-classified sentences are available online along with our set of errorful sentences.

[3] Data is available at `http://pages.cs.wisc.edu/~bsnyder`

in Table 1. Our model does not allow reorderings, and can thus be described in terms of word-by-word stochastic transformations to the adult sentence.

We use 10 word classes to parameterize our model: pronouns, negators, wh-words, conjunctions, prepositions, determiners, modal verbs, "be" verbs, other auxiliary verbs, and lexical content words. The list of words in each class is provided as part of our data set. For each input adult word w, the model generates output word w' as a hierarchical series of draws from multinomial distributions, conditioned on the original word w and its class c.

All distributions receive an asymmetric Dirichlet prior which favors retention of the adult word. With the sole exception of word insertions, the distributions are parameterized and learned during training. Our model consists of 217 multinomial distributions, with 6,718 free parameters.

The precise form and parameterization of our model were handcrafted for performance on the development data, using trial and error. We also considered more fine-grained model forms (i.e. one parameter for each non-lexical input-output word pair), as well as coarser parameterizations (i.e. a single shared parameter denoting any inflection change). The model we describe here seemed to achieve the best balance of specificity and generalization.

We now present pseudocode describing the noise model's operation upon processing each word, along with a brief description of each step.

Action selection (lines 3-7): On reading an input word, an action category a is selected from a probability distribution conditioned on the input word's class. Our model allows up to two function word insertions or deletions in a row before a swap is required. Lexical content words may not be deleted or inserted, only swapped.

Insert and Delete (lines 8-15): The deletion case requires no decision after action selection. In the insertion case, the class of the inserted word, c', is selected conditioned on c_{PREV}, the class of the previous adult word. The precise identity of the inserted word is then drawn from a uniform distribution over words in class c'. It is important to note that in the insertion case, the input word at a given iteration will be re-processed at the next iteration (lines 33-35).

```
    insdel ← 0
    for word w with class c, inflection f, lemma ℓ
    do
 3:   if insdel = 2 then
          a ← swap
      else
 6:       a ~ {insert, delete, swap} | c
      end if
      if a = delete then
 9:       insdel++
          c' ← ϵ
          w' ← ϵ
12:   else if a = insert then
          insdel++
          c' ~ classes | c_PREV, insert
15:       w' ~ words in c' | insert
      else
          insdel ← 0
18:       c' ← c
          if c ∈ uninflected-classes then
              w' ~ words in c | w, swap
21:       else if c = aux then
              ℓ' ~ aux-lemmas | ℓ, swap
              f' ~ inflections | f, swap
24:           w' ← COMBINE(ℓ', f')
          else
              f' ~ inflections | f, swap
27:           w' ← COMBINE(ℓ, f')
          end if
      end if
30:   if w' ∈ irregular then
          w' ~ OVERGEN(w') ∪ {w'}
      end if
33:   if a = insert then
          goto line 3
      end if
36: end for
```

Swap (lines 16 - 29): In the swap case, a word of given class is substituted for another word in the same class. Depending on the source word's class, swaps are handled in slightly different ways. If the word is a modal, conjunction, determiner, preposition, "wh-" word or negative, it is considered "unin-

flected." In these cases, a new word w' is selected from all words in class c, conditioned on the source word w.

If w is an auxiliary verb, the swap procedure consists of two parallel steps. A lemma is selected from possible auxiliary lemmas, conditioned on the lemma of the source word.[4] In the second step, an output inflection type is selected from a distribution conditioned on the source word's inflection. The precise output word is fully specified by the choice of lemma and conjugation.

If w is not in either of the above two categories, it is a lexical word, and our model only allows changes in conjugation or declension. If the source word is a noun it may swap to singular or plural form conditioned on the source form. If the word is a verb, it may swap to any conjugated or non-finite form, again conditioned on the source form. Lexical words that are not marked by CELEX (Baayen et al., 1996) as nouns or verbs may only swap to the exact same word.

Overgeneralization (lines 30-32): Finally, the noisy channel considers the possibility of producing overgeneralized word forms (like "maked" and "childs") in place of their correct irregular forms. The OVERGEN function produces the incorrect overgeneralized form. We draw from a distribution which chooses between this form and the correct original word. Our model maintains separate distributions for nouns (overgeneralized plurals) and verbs (overgeneralized past tense).

5 Implementation

In this section, we describe steps necessary to build, train and test our error correction model. Weighted Finite State Transducers (FSTs) used in our model are constructed with OpenFst (Allauzen et al., 2007).

5.1 Sentence FSTs

These FSTs provide the basis for our translation process. We represent sentences by building a simple linear chain FST, progressing from node to node with each arc accepting and yielding one word in the sentence. All arcs are weighted with probability one.

[4] Auxiliary lemmas include *have, do, go, will,* and *get*.

5.2 Noise FST

The noise model provides a conditional probability over child sentences given an adult sentence. We encode this model as a FST with several states, allowing us to track the number of consecutive insertions or deletions. We allow only two of these operations in a row, thereby constraining the length of the output sentence. This constraint results in three states ($insdel = 0, insdel = 1, insdel = 2$), along with an end state. In our training data, only 2 sentence pairs cannot be described by the noise model due to this constraint.

Each arc in the FST has an ϵ or adult-language word as input symbol, and a possibly errorful child-language word or ϵ as output symbol. Each arc weight is the probability of transducing the input word to the output word, determined according to the parameterized distributions described in Section 4. Arcs corresponding to insertions or deletions lead to a new state ($insdel$++) and are not allowed from state $insdel = 2$. Substitution arcs all lead back to state $insdel = 0$. Word class information is given by a set of word lists for each non-lexical class.[5] Inflectional information is derived from CELEX.

5.3 Language Model FST

The language model provides a prior distribution over adult form sentences. We build a a trigram language model FST with Kneser-Ney smoothing using OpenGRM (Roark et al., 2012). The language model is trained on all parent speech in the CHILDES studies from which our errorful sentences are drawn.

In the language model FST, the input and output words of each arc are identical. Arcs are weighted with the probability of the n-gram beginning with some prefix associated with the source node, and ending with the arc's input/output word. In this setup, the probability of a string is the total weight of the path accepting and emitting that string.

5.4 Training

As detailed in Section 4, our noise model consists of a series of multinomial distributions which govern

[5] Word lists are included for reference with our dataset.

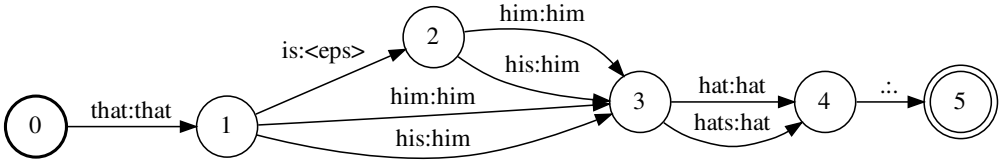

Figure 1: A simplified decoding FST for the child sentence "That him hat." In an actual decoding FST many more transduction arcs exist, including those translating "that" and "him" to any determiner and pronoun, respectively, and affording opportunities for many more deletions and insertions. Input and output strings given by FST paths correspond to possible adult-to-child translations.

the transformation from adult word to child word, allowing limited insertions and deletions. We estimate parameters θ for these distributions that maximize their posterior probability given the observed training sentences $\{(s,t)\}$. Since our language model $P(t)$ does not depend on on the noise model parameters, this objective is equivalent to jointly maximizing the prior and the conditional likelihoods of child sentences given adult sentences:

$$\operatorname*{argmax}_{\theta} P(\theta) \prod P(s|t,\theta)$$

To represent all possible derivations of each child sentence s from its adult translation t, we compose the sentence FSTs with the noise model, obtaining:

$$FST_{train} = FST_t \circ FST_{noise} \circ FST_s$$

Each path through FST_{train} corresponds to a single derivation d, with path weight $P(s,d|t,\theta)$. By summing all path weights, we obtain $P(s|t,\theta)$. We use a MAP-EM algorithm to maximize our objective while summing over all possible derivations.

Our training scheme relies on FSTs weighted in the V-expectation semiring (Eisner, 2001), implemented using code from fstrain (Dreyer et al., 2008). Besides carrying probabilities, arc weights are supplemented with a vector to indicate parameter counts involved in the arc traversal. The V-expectation semiring is designed so that the total arc weight of all paths through the FST yields both the probability $P(s|t,\theta)$, along with expected parameter counts. Our EM algorithm proceeds as follows: We start by initializing all parameters to uniform distributions with random noise. We then weight the arcs in FST_{noise} accordingly. For each sentence pair (s,t), we build FST_{train} by composition with our noise model, as described in the previous paragraph. We then compute the total arc weight of all paths through FST_{train} by relabeling all input and output symbols to ϵ and then reducing FST_{train} to a single state using epsilon removal (Mohri, 2008). The stopping weight of this single state is the sum of all paths through the original FST, yielding the probability $P(s|t,\theta)$, along with expected parameter counts according to our current distributions. We then reestimate θ using the expected counts plus pseudo-counts given by priors, and repeat this process until convergence.

Besides smoothing our estimated distributions, the pseudo-counts given by our asymmetric Dirichlet priors favor multinomials that retain the adult word form (swaps, identical lemmas, and identical inflections). Concretely, we use pseudo-counts of .5 for these favored outcomes, and pseudo-counts of .01 for all others.[6]

In practice, 109 of the child sentences in our data set cannot be translated into a corresponding adult version using our model. This is due to a range of rare phenomena like rephrasing, lexical word swaps and word-order errors. In these cases, the composed FST has no valid paths from start to finish and the sentence is removed from training. We run EM for 100 iterations, at which time the log likelihood of all sentences generally converges to within .01.

5.5 Decoding

After training our noise model, we apply the system to translate divergent child language to adult-like speech. As in training, the noise FST is composed with the FST for each child sentence s. In

[6] corresponding to Dirichlet hyperparameters of 1.5 and 1.01 respectively.

place of the adult sentence, the language model FST is used, yielding:

$$FST_{decode} = FST_{lm} \circ FST_{noise} \circ FST_s$$

Each path through FST_{decode} corresponds to an adult translation and derivation (t, d), with path weight $P(s, d|t, \theta)P(t)$. Thus, the highest-weight path corresponds to the most likely translation and derivation pair:

$$\underset{t,d}{\operatorname{argmax}} P(t, d|s, \theta)$$

We use a dynamic program to find the n highest-weight paths with distinct adult sentences t. This can be viewed as finding the n most likely adult translations, using a Viterbi approximation $P(t|s, \theta) = \operatorname{argmax}_d P(t, d|s, \theta)$. In our experiments we set $n = 50$. A simplified FST_{decode} example is shown in Figure 1.

5.6 Discriminative Reranking

To more flexibly capture long range syntactic features, we embed our noisy channel model in a discriminative reranking procedure. For each child sentence s, we take the n-best candidate translations $t_1, \ldots, t_n$ from the underlying generative model, as described in the previous section. We then map each candidate translation t_i to a d-dimensional feature vector $f(s, t_i)$. The reranking model then uses a d-dimensional weight vector λ to predict the candidate translation with highest linear score:

$$t^* = \underset{t_i}{\operatorname{argmax}} \lambda \cdot f(s, t_i)$$

To simulate test conditions, we train the weight vector on n-best lists from 8-fold cross-validation over training data, using the averaged perceptron reranking algorithm (Collins and Roark, 2004). Since the n-best list might not include the exact gold-standard correction, a target correction which maximizes our evaluation metric is chosen from the list. The n-best list is non-linearly separable, so perceptron training iterates for 1000 rounds, when it is terminated without converging.

Our feature function $f(s, t_i)$ yields nine boolean and real-valued features derived from (i) the FST that generates child sentence s from candidate adult-form t_i, and (ii) the POS sequence and dependency parse of candidate t_i obtained with the Stanford Parser (de Marneffe et al., 2006). Features were selected based on their performance in reranking held-out development data from the training set. Reranking features are given below:

Generative Model Probabilities: We first include the joint probability of the child sentence s and candidate translation t_i, given by the generative model: $P_{lm}(t_i)P_{noise}(s|t_i)$. We also isolate the candidate translation's language model and noise model probabilities as features. Since both of these probabilities naturally favor shorter sentences, we scale them to sentence length, yielding $\sqrt[n]{P_{lm}(t_i)}$ and $\sqrt[n]{P_{noise}(s|t_i)}$ respectively. By not scaling the joint probability, we allow the reranker to learn its own bias towards longer or shorter corrected sentences.

Contains Noun Subject, Accusative Noun Subject: The first boolean feature indicates whether the dependency parse of candidate translation t_i contains a "nsubj" relation. The second indicates if a "nsubj" relation exists where the dependent is an accusative pronoun (e.g. "Him ate the cookie"). These features and the one following have previously been used in classifier based error detection (Morley and Prud'hommeaux, 2012).

Contains Finite Verb: This boolean feature is true if the POS tags of t_i include a finite verb. This feature differentiates structures like "I am going" from "I going."

Question Template Features: We define templates for wh- and yes-no questions. A sentence fits the wh- question template if it begins with a wh-word, followed by an auxiliary or copula verb (e.g. "Who did..."). A sentence fits the yes-no template when it begins with an auxiliary or copula verb, then a noun subject followed by a verb or adjective (e.g. "Are you going..."). We include one boolean feature for each of these templates indicating when a template match is *inappropriate*, when the original child utterance terminates in a period instead of a question mark. In addition to the two features for inappropriate template matches, we have a single feature that signals *appropriate* matches of either question template — when the original child utterance terminates in a question mark.

Child Utterance	Human Correction	Machine Correction
I am not put in my mouth.	I am not putting it in my mouth.	I am not going to put it in my mouth.
This one have water?	Does this one have water?	This one has water?
Want to read the book.	I want to read the book.	You want to read the book.
Why you going to get two?	Why are you going to get two?	Why are you going to have two?
You very sticky.	You are very sticky.	You are very sticky.
He no like.	He does not like it.	He does not like that.
Yeah it looks a lady.	Yeah it looks like a lady	Yeah it looks like a lady.
Eleanor come too.	Eleanor came too.	Eleanor come too.
Desk in here.	The desk is in here	Desk is in here.
Why he's doc?	Why is he called doc?	He's up doc?

Table 2: Randomly selected test output generated by our complete error correction model, along with corresponding child utterances and human corrections.

6 Experiments and Analysis

Baselines We compare our system's performance with two pre-existing baselines. The first is a standard phrase-based machine translation system using MOSES (Koehn et al., 2007) with GIZA++ (Och and Ney, 2003) word alignments. We hold out 9% of the training data for tuning using the MERT algorithm with BLEU objective (Och, 2003).

The second baseline is our implementation of the ESL error correction system described by Park and Levy (2011). Like our system, this baseline trains FST noise models using EM in the V-expectation semiring. Our noise model is crafted specifically for the child language domain, and so differs from Park and Levy's in several ways: First, we capture a wider range of word-swaps, with richer parameterization allowing many more translation options. As a result, our model has 6,718 parameters, many more than the ESL model's 187. These parameters correspond to learned probability distributions, whereas in the ESL model many of the distributions are fixed as uniform. We also capture a larger class of errors, including deletions, change of auxiliary lemma, and inflectional overgeneralizations. Finally, we use a discriminative reranking step to model long-range syntactic dependencies. Although the ESL model is originally geared towards fully unsupervised training, we train this baseline in the same supervised framework as our model.

Evaluation and Performance We train all models on 80% of our child-adult sentence pairs and test on the remaining 20%. For illustration, selected output from our model is shown in Table 2.

Predictions are evaluated with BLEU score (Papineni et al., 2002) and Word Error Rate (WER), defined as the minimum string edit distance (in words) between reference and predicted translations, divided by length of the reference. As a control, we compare all results against scores for the uncorrected child sentences themselves. As reported in Table 3, our model achieves the best scores for both metrics. BLEU score increases from 50 for child sentences to 62, while WER is reduced from .271 to .224. Interestingly, MOSES achieves a BLEU score of 58 — still four points below our model — but actually increases WER to .449. For both metrics, the ESL system increases error. This is not surprising given that its intended application is in an entirely different domain.

Error Analysis We measured the performance of our model over the six most common categories of child divergence, including deletions of various function words and overgeneralizations of past tense forms (e.g. "maked" for "made"). We first identified model parameters associated with each category, and then counted the number of correct and incorrect parameter firings on the test sentences. As Table 4 indicates, our model performs reasonably well on "be" verb deletions, preposition deletions, and overgeneralizations, but has difficulty correcting pronoun and auxiliary deletions.

In general, hypothesizing dropped words burdens the noise model by adding additional draws from multinomial distributions to the derivation. To pre-

	BLEU	WER
WER reranking	**62.12**	**.224**
BLEU reranking	60.86	.231
No reranking	60.37	.233
Moses	58.29	.449
ESL	40.76	.318
Child Sentences	49.55	.271

Table 3: WER and BLEU scores. Our system's performance using various reranking schemes (BLEU objective, WER objective and none) is contrasted with Moses MT and ESL error correction baselines, as well as uncorrected test sentences. Best performance under each metric is shown in bold.

dict a deletion, either the language model or the reranker must strongly prefer including the omitted word. A syntax-based noise model may achieve better performance in detecting and correcting child word drops.

While our model parameterization and performance rely on the largely constrained nature of child language errors, we observe some instances in which it is overly restrictive. For 10% of utterances in our corpus, it is impossible to recover the exact gold-standard adult sentence. These sentences feature errors like reordering or lexical lemma swaps — for example "I talk Mexican" for "I speak Spanish." While our model may correct other errors in these sentences, a perfect correction is unattainable.

Sometimes, our model produces appropriate forms which by happenstance do not conform to the annotators' decision. For example, in the second row of Table 2, the model corrects "This one have water?" to "This one has water?", instead of the more verbose correction chosen by the annotators ("Does this one have water?"). Similarly, our model sometimes produces corrections which seem appropriate in isolation, but do not preserve the meaning implied by the larger conversational context. For example, in row three of Table 2, the sentence "Want to read the book." is recognized both by our human annotators and the system as requiring a pronoun subject. Unlike the annotators, however, the model has no knowledge of conversational context, so it chooses the highest probability pronoun — in this case "you" — instead of the contextually correct "I."

Error Type	Count	F_1	P	R
Be Deletions	63	.84	.84	.84
Pronoun Deletions	30	.15	.38	.1
Aux. Deletions	30	.21	.44	.13
Prep. Deletions	26	.65	.82	.54
Det. Deletions	22	.48	.73	.36
Overgen. Past	7	.92	1.0	.86

Table 4: Frequency of the six most common error types in test data, along with our model's corresponding F-measure, precision and recall. All counts are $\pm.12$ at $p = .05$ under a binomial normal approximation interval.

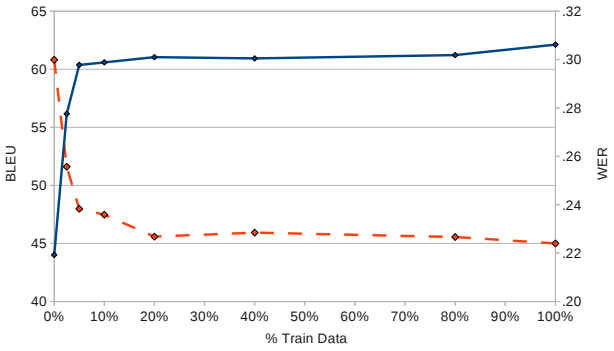

Figure 2: Performance with limited training data. WER is drawn as the dashed line, and BLEU as the solid line.

Learning Curves In Figure 2, we see that the learning curves for our model initially rise sharply, then remain relatively flat. Using only 10% of our training data (80 sentences), we increase BLEU from 44 (using just the language model) to almost 61. We only reach our reported BLEU score of 62 when adding the final 20% of training data. This result emphasizes the specificity of our parameterization. Because our model is so tailored to the child-language scenario, only a few examples of each error type are needed to find good parameter values. We suspect that more annotated data would lead to a continued but slow increase in performance.

Training and Testing across Children We use our system to investigate the hypothesis that language acquisition follows a similar path across children (Brown, 1973). To test this hypothesis, we train our model on all children excluding Adam, who alone is responsible for 21% of our sentences. We then test the learned model on the separated Adam

Trained on:	**BLEU**	**WER**
Adam	**72.58**	.226
All Others	69.83	**.186**
Uncorrected	45.54	.278

Table 5: Performance on Adam's sentences training on other children, versus training on himself. Best performance under each metric is shown in bold.

data. These results are contrasted with performance of 8-fold cross validation training and testing solely on Adam's utterances. Performance statistics are given in Table 5.

We first note that models trained in both scenarios lead to large error reductions over the child sentences. This provides evidence that our model captures general, and not child-specific, error patterns. Although training exclusively on Adam does lead to increased BLEU score (72.58 vs 69.83), WER is minimized when using the larger volume of training data from other children (.186 vs .226). Taken as a whole, these results suggest that training and testing on separate children does not degrade performance. This finding supports the general hypothesis of shared developmental paths.

Plotting Child Language Errors over Time After training on annotated data, we predict divergences in all available data from the children in Roger Brown's 1973 study — Adam, Eve and Sarah — as well as Abe (Kuczaj, 1977), a child from a separate study over a similar age-range. We plot each child's per-utterance frequency of preposition omissions in Figure 3. Since we evaluate over 65,000 utterances and reranking has no impact on preposition drop prediction, we skip the reranking step to save computation.

In Figure 3, we see that Adam and Sarah's preposition drops spike early, and then gradually decrease in frequency as their preposition use moves towards that of an adult. Although Eve's data covers an earlier time period, we see that her pattern of preposition drops shows a similar spike and gradual decrease. This is consistent with Eve's general language precocity. Brown's conclusion — that the language development of these three children advanced in similar stages at different times — is consistent with our predictions. However, when we examine

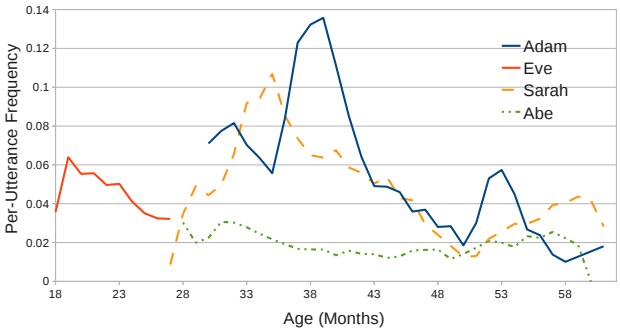

Figure 3: Automatically detected preposition omissions in un-annotated utterances from four children over time. Assuming perfect model predictions, frequencies are $\pm.002$ at $p = .05$ under a binomial normal approximation interval. Prediction error is given in Table 4.

Abe we do not observe the same pattern.[7] This points to a degree of variance across children, and suggests the use of our model as a tool for further empirical refinement of language development hypotheses.

Discussion Our error correction system is designed to be more constrained than a full-scale MT system, focusing parameter learning on errors that are known to be common to child language learners. Reorderings are prohibited, lexical word swaps are limited to inflectional changes, and deletions are restricted to function word categories. By highly restricting our hypothesis space, we provide an inductive bias for our model that matches the child language domain. This is particularly important since the size of our training set is much smaller than that usually used in MT. Indeed, as Figure 2 shows, very little data is needed to achieve good performance.

In contrast, the ESL baseline suffers because its generative model is too restricted for the domain of transcribed child language. As shown above in Table 4, child deletions of function words are the most frequent error types in our data. Since the ESL model does not capture word deletions, and has a more restricted notion of word swaps, 88% of child sentences in our training corpus cannot be translated to their reference adult versions. The result is that the ESL model tends to rely too heavily on the language model. For example, on the sentence "I com-

[7] Though it is of course possible that a similar spike and drop-off occurred earlier in Abe's development.

ing to you," the ESL model improves n-gram probability by producing "I came to you" instead of the correct "I am coming to you". This increases error over the child sentence itself.

In addition to the domain-specific generative model, our approach has the advantage of long-range syntactic information encoded by reranking features. Although the perceptron algorithm places high weight on the generative model probability, it alters the predictions in 17 out of 201 test sentences, in all cases an improvement. Three of these reranking changes add a noun subject, five enforce question structure, and nine add a main verb.

7 Conclusion and Future Work

In this paper we introduce a corpus of divergent child sentences with corresponding adult forms, enabling the systematic computational modeling of child language by relating it to adult grammar. We propose a child-to-adult translation task as a means to investigate child language development, and provide an initial model for this task.

Our model is based on a noisy-channel assumption, allowing for the deletion and corruption of individual words, and is trained using FST techniques. Despite the debatable cognitive plausibility of our setup, our results demonstrate that our model captures many standard divergences and reduces the average error of child sentences by approximately 20%, with high performance on specific frequently occurring error types.

The model allows us to chart aspects of language development over time, without the need for additional human annotation. Our experiments show that children share common developmental stages in language learning, while pointing to child-specific subtleties in preposition use.

In future work, we intend to dynamically model child language ability as it grows and shifts in response to internal processes and external stimuli. We also plan to develop and train models specializing in the detection of specific error categories. By explicitly shifting our model's objective from child-adult translation to the detection of some particular error, we hope to improve our analysis of child divergences over time.

Acknowledgments

The authors thank the reviewers and acknowledge support by the NSF (grant IIS-1116676) and a research gift from Google. Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the NSF.

References

A. Alishahi. 2010. Computational modeling of human language acquisition. *Synthesis Lectures on Human Language Technologies*, 3(1):1–107.

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. *Implementation and Application of Automata*, pages 11–23.

R.H. Baayen, R. Piepenbrock, and L. Gulikers. 1996. CELEX2 (CD-ROM). Linguistic Data Consortium.

E. Bates, I. Bretherton, and L. Snyder. 1988. *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge University Press.

D.C. Bellinger and J.B. Gleason. 1982. Sex differences in parental directives to young children. *Sex Roles*, 8(11):1123–1139.

L. Bliss. 1988. The development of modals. *Journal of Applied Developmental Psychology*, 9:253–261.

L. Bloom, L. Hood, and P. Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive Psychology*, 6(3):380–420.

L. Bloom, P. Lightbown, L. Hood, M. Bowerman, M. Maratsos, and M.P. Maratsos. 1975. Structure and variation in child language. *Monographs of the Society for Research in Child Development*, pages 1–97.

L. Bloom. 1973. *One word at a time: The use of single word utterances before syntax*. Mouton.

P. Bloom. 1990. Subjectless sentences in child language. *Linguistic Inquiry*, 21(4):491–504.

J.N. Bohannon III and A.L. Marquis. 1977. Children's control of adult speech. *Child Development*, 48(3):1002–1008.

R. Brown. 1973. *A first language: The early stages.* Harvard University Press.

V. Carlson-Luden. 1979. *Causal understanding in the 10-month-old*. Ph.D. thesis, University of Colorado at Boulder.

E.C. Carterette and M.H. Jones. 1974. *Informal speech: Alphabetic & phonemic texts with statistical analyses and tables*. University of California Press.

M. Chodorow and C. Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 140–147.

M. Collins and B. Roark. 2004. Incremental parsing with the perceptron algorithm. In *Proceedings of the Association for Computational Linguistics*, pages 111–118, Barcelona, Spain, July.

M. Connor, Y. Gertner, C. Fisher, and D. Roth. 2008. Baby SRL: Modeling early language acquisition. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 81–88.

R. Dale and A. Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the European Workshop on Natural Language Generation*, pages 242–249.

M.C. de Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of The International Conference on Language Resources and Evaluation*, volume 6, pages 449–454.

M.J. Demetras, K.N. Post, and C.E. Snow. 1986. Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language*, 13(2):275–292.

M.J. Demetras. 1989. Working parents' conversational responses to their two-year-old sons.

M. Dreyer, J.R. Smith, and J. Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1080–1089.

A. Echihabi and D. Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the Association for Computational Linguistics*, pages 16–23.

J. Eisner. 2001. Expectation semirings: Flexible EM for learning finite-state transducers. In *Proceedings of the ESSLLI workshop on finite-state methods in NLP*.

M. Gamon. 2011. High-order sequence modeling for language learner error detection. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 180–189.

L.C.G. Haggerty. 1930. What a two-and-one-half-year-old child said in one day. *The Pedagogical Seminary and Journal of Genetic Psychology*, 37(1):75–101.

W.S. Hall, W.C. Tirre, A.L. Brown, J.C. Campoine, P.F. Nardulli, HO Abdulrahman, MA Sozen, W.C. Schnobrich, H. Cecen, J.G. Barnitz, et al. 1979. The communicative environment of young children: Social class, ethnic, and situational differences. *Bulletin of the Center for Children's Books*, 32:08.

W.S. Hall, W.E. Nagy, and R.L. Linn. 1980. *Spoken words: Effects of situation and social group on oral word usage and frequency*. University of Illinois at Urbana-Champaign, Center for the Study of Reading.

W.S. Hall, W.E. Nagy, and G. Nottenburg. 1981. Situational variation in the use of internal state words. Technical report, University of Illinois at Urbana-Champaign, Center for the Study of Reading.

H. Hamburger and S. Crain. 1984. Acquisition of cognitive compiling. *Cognition*, 17(2):85–136.

R.P. Higginson. 1987. *Fixing: Assimilation in language acquisition*. University Microfilms International.

M.H. Jones and E.C. Carterette. 1963. Redundancy in children's free-reading choices. *Journal of Verbal Learning and Verbal Behavior*, 2(5-6):489–493.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (Interactive Poster and Demonstration Sessions)*, pages 177–180.

S. A. Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

J. Lee and S. Seneff. 2006. Automatic grammar correction for second-language learners. In *Proceedings of the International Conference on Spoken Language Processing*.

X. Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, volume 2. Psychology Press.

B. MacWhinney. 2007. The TalkBank project. *Creating and digitizing language corpora: Synchronic Databases*, 1:163–180.

M. Mohri. 2008. System and method of epsilon removal of weighted automata and transducers, June 3. US Patent 7,383,185.

E. Morley and E. Prud'hommeaux. 2012. Using constituency and dependency parse features to identify errorful words in disordered language. In *Proceedings of the Workshop on Child, Computer and Interaction*.

A. Ninio, C.E. Snow, B.A. Pan, and P.R. Rollins. 1994. Classifying communicative acts in children's interactions. *Journal of Communication Disorders*, 27(2):157–187.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 160–167.

R.E. Owens. 2008. *Language development: An introduction*. Pearson Education, Inc.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Y.A. Park and R. Levy. 2011. Automated whole sentence grammar correction using a noisy channel model. *Proceedings of the Association for Computational Linguistics*, pages 934–944.

A.M. Peters. 1987. The role of imitation in the developing syntax of a blind child in perspectives on repetition. *Text*, 7(3):289–311.

K. Post. 1992. *The language learning environment of laterborns in a rural Florida community*. Ph.D. thesis, Harvard University.

C. Quirk, C. Brockett, and W. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 142–149.

T. Regier. 2005. The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29(6):819–865.

A. Ritter, C. Cherry, and W.B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, and T. Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the Association for Computational Linguistics (System Demonstrations)*, pages 61–66.

A. Rozovskaya, M. Sammons, J. Gioja, and D. Roth. 2011. University of Illinois system in HOO text correction shared task. In *Proceedings of the European Workshop on Natural Language Generation*, pages 263–266.

J. Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language*, 4.

K. Sagae, A. Lavie, and B. MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the Association for Computational Linguistics*, pages 197–204.

S. Sahakian and B. Snyder. 2012. Automatically learning measures of child language development. *Proceedings of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 95–99.

C.E. Snow, F. Shonkoff, K. Lee, and H. Levin. 1986. Learning to play doctor: Effects of sex, age, and experience in hospital. *Discourse Processes*, 9(4):461–473.

E.L. Stine and J.N. Bohannon. 1983. Imitations, interactions, and language acquisition. *Journal of Child Language*, 10(03):589–603.

X. Sun, J. Gao, D. Micol, and C. Quirk. 2010. Learning phrase-based spelling error models from clickthrough data. In *Proceedings of the Association for Computational Linguistics*, pages 266–274.

P. Suppes. 1974. The semantics of children's language. *American Psychologist*, 29(2):103.

T.Z. Tardif. 1994. *Adult-to-child speech and language acquisition in Mandarin Chinese*. Ph.D. thesis, Yale University.

V. Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition*, 40(1-2):21–81.

L. Van Houten. 1986. Role of maternal input in the acquisition process: The communicative strategies of adolescent and older mothers with their language learning children. In *Boston University Conference on Language Development*.

A. Warren-Leubecker and J.N. Bohannon III. 1984. Intonation patterns in child-directed speech: Mother-father differences. *Child Development*, 55(4):1379–1385.

A. Warren. 1982. *Sex differences in speech to children*. Ph.D. thesis, Georgia Institute of Technology.

B. Wilson and A.M. Peters. 1988. What are you cookin' on a hot?: A three-year-old blind child's 'violation' of universal constraints on constituent movement. *Language*, 64:249–273.