Contextual Bootstrapping for Grammar Learning

Eva H. Mok

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-179
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-179.html

December 19, 2008

Contextual Bootstrapping for Grammar Learning

by

Eva H. Mok

B.S. (University of Michigan, Ann Arbor) 2000

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Jerome A. Feldman, Chair
Professor Carla Hudson Kam
Professor Dan Klein

Fall 2008

The dissertation of Eva H. Mok is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2008

Contextual Bootstrapping for Grammar Learning

Abstract

Contextual Bootstrapping for Grammar Learning

by

Eva H. Mok

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Jerome A. Feldman, Chair

The problem of grammar learning is a challenging one for both children and machines due to impoverished input: hidden grammatical structures, lack of explicit correction, and in pro-drop languages, argument omission. This dissertation describes a computational model of child grammar learning using a probabilistic version of *Embodied Construction Grammar* (ECG) that demonstrates how the problem of impoverished input is alleviated through bootstrapping from the situational context. This model represents the convergence of: (1) a unified representation that integrates semantic knowledge, linguistic knowledge, and contextual knowledge, (2) a context-aware language understanding process, and (3) a structured grammar learning and generalization process.

Using situated child-directed utterances as learning input, the model performs two concurrent learning tasks: structural learning of the grammatical units and statistical learning of the associated parameters. The structural learning task is a guided search over the space of possible constructions. The search is informed by embodied semantic knowledge that it has gathered through experience with the world even before learning grammar and situational knowledge that the model obtains from context. The statistical learning task requires continuous updating of the parameters associated with the probabilistic grammar based on usage and these parameters reflect shifting preferences on learned grammatical structures.

The computational model of grammar learning has been validated in two ways. It has been applied to a subset of the CHILDES Beijing corpus, which is a corpus of naturalistic parent-child interaction in Mandarin Chinese. Its learning behavior has also been more closely examined using an artificial miniature language. This learning model provides a precise, computational framework for fleshing out theories of construction formation and generalization.

Professor Jerome A. Feldman
Dissertation Committee Chair

In memory of my father

Contents

Acknowledgements

I am very fortunate to have been surrounded by extremely supportive friends over the years. I could never have made it this far without them, and I can never thank them enough.

To my advisor Jerry, who has given me the freedom to do the research that I love.

To my secondary advisor Carla, who has guided me into the field of language development, and all her former and current lab members, who graciously took me in.

To my tertiary advisor Srini, who has always been willing to spend an hour to discuss research with me when he only has a minute.

To my committee member Dan, who has always given me very insightful feedback.

To ICSI and all the folks there, who have provided excellent infrastructural support.

To my friend and officemate Nancy, who not only got me started on computational modeling research, but has also been helpful and supportive in more ways than I can describe.

To my awesome friend and colleague Johno, who has been very fun to collaborate with.

To my dear friend and colleague Steve, who saw me through all the difficult times.

To my life-long friends Jeff and Bryan, who have cheered me on and kept me motivated by periodically asking the dreaded question, "Are you done yet?"

To my friends Joe, Leon, Michael, Josef, Ellen, Ben, Palani, Selden, and Mariianne, who have kept me in good spirits as I struggled through this dissertation.

And of course, to my mom and dad.

Chapter 1.

Modeling the Learning of Contextual Constructions

The level of competence that children achieve in their native language in a bare four to five years is a remarkable feat given the intricacy and nuances of language. The key problem in language acquisition is that the linguistic input alone vastly underdetermines the hidden structures that are generally attributed as grammatical knowledge, and this problem is pervasive in every aspect of language from phonology to pragmatics. At the phonological level, word segments are not denoted in fluent speech by pauses (Jusczyk, 1997) and a child must learn to pick out the words. The task of word learning is plagued with the problem of indeterminacy (such Quine's famous "gavagai" example (1960)), which is arguably worse for verbs than for nouns. In grammar learning, syntactic structures are not at all present in the input, and yet the ability to productively manipulate these structures is considered to be the defining characteristic of grammatical knowledge. It seems miraculous that normally developing children become such competent language users in such a short period of time.

Indeed, the very complexity in the task of language learning has been used to argue for the innateness of language, and the acquisition of syntax is at the heart of this debate. There are good theoretical and psychological reasons, however, to believe that innate knowledge of linguistic principles and parameters need not be the case. In contrast to Gold's theorem (1967), which shows that categorical regular languages and context-free languages are not identifiable in the limit on the basis of positive examples alone, Horning (1969) demonstrates that stochastic context-free grammars are learnable given some assumptions about the priors of the grammars. Many counterarguments to innateness have also been offered on the psychological end of the debate, from work directly addressing the poverty of stimulus claim by looking at the input children receive (Pullum & Scholz, 2002) to work addressing the logical problem of language acquisition by offering alternative mechanistic accounts (Macwhinney, 2004; Perfors, 2008; Perfors, Tenenbaum & Regier, 2006). In addition, there are calls for an alternate conceptualization of the innateness debate that studies the interaction between genetics and environment (1997). Instead of rehashing old arguments, this dissertation takes as a starting point the assumption that language is too complex a system to be learned through blind associations between linguistic and non-linguistic input. Some form of learning bias must be introduced into the learning process; it is the goal of this dissertation to lay out systematically, in a computational framework, some learning biases that facilitate the process without resorting to innate knowledge of syntax.

Put in concrete terms, this dissertation is concerned with modeling how semantic knowledge about typical actions and events and contextual knowledge about the situation surrounding each piece of learning input come together to aid the acquisition of grammar, or more precisely, the language-specific ways in which relational meanings between words are denoted. For example, the semantic relation that John is the hitter and the ball is the hittee of a hitting event is denoted by word order in the English sentence *John hit the ball*. Grammar, in this formulation, consists not of syntactic rules that allow or disallow sentences in a language but of conventionalized mappings between linguistic forms and embodied meanings. Building on the construction grammar framework (Fillmore et al. 1988; Goldberg 1995; Fillmore et al. 1999; Kay et al. 1999), grammatical knowledge comprises form-meaning mappings as rigid as idioms (e.g.

cross your fingers) or early holophrases (e.g. *gimme that*) and as broad and productive as the ditransitive / double-object construction (e.g. *John baked Mary a cake*).

To begin, it is important to keep in mind that as a child receives language input throughout infancy, she also accumulates experience with the world through physical interaction with objects and communicative acts with her caregivers. By the time she starts attending to syntactic cues at around 17 months of age (Hirsh-Pasek, Golinkoff & Naigles, 1996a), she has access to a repertoire of embodied knowledge about various objects, people, and motor actions that form a rich substrate for language learning. Furthermore, a child learns language not in isolation but in context structured by rituals and routines which help guide the child's interpretation of novel utterances.

The support that context lends to language comprehension is especially pronounced in pronoun-dropping (pro-drop) languages, where it is not only permissible but common to omit the subject and/or object from sentences (also referred to as zero anaphora or null subject/object). In some languages subject omission is the most prevalent (e.g. Spanish, Italian); in others both subject and object omissions are permissible (e.g. Chinese, Korean). Furthermore, unlike some morphologically rich pro-drop languages, a language like Mandarin Chinese has little inflectional morphology that helps to constrain the interpretation of the omitted referent. In this case, understanding an utterance requires not just ongoing awareness of the situational and discourse context but also inference mechanisms to arrive at the most plausible reference interpretation. This is the notion of best-fit constructional analysis as described in (Bryant, 2008a).

To give the reader a better idea of what this best-fit constructional analysis process entails, Figure 1.1 shows the interpretation of a sentence in Mandarin Chinese (which is a SVO language), *xi1xi1 chi1 yao4* (XiXi eat medicine). A hearer who knows how Mandarin works realizes that this is a transitive sentence and the word order in the sentence signifies that XiXi the child is the eater and some Medicine is the eatee of an EAT action denoted by the word *chi1*. In the figure, these word order and meaning relations are signified with bold solid arrows on the left. Language understanding is much more than thematic role assignment, however: the hearer has access to embodied knowledge about eating and knows that eating typically involves chewing and swallowing as sub-processes. She is able to immediately realize that the eater (in this case the child) is also the chewer and the swallower if she needs to reason about the eating event.

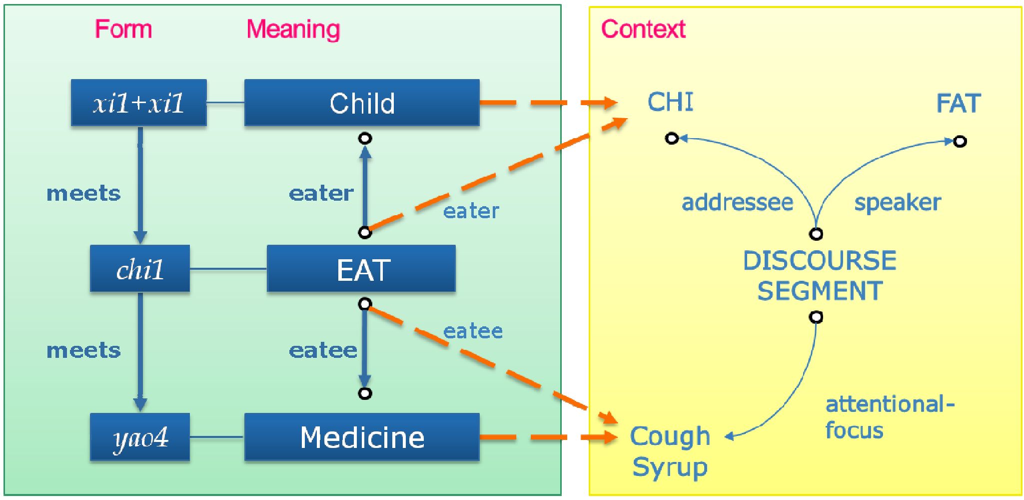

Figure 1.1 Interpreting the sentence *xi1xi1 chi1 yao4* (Xixi eat medicine) involves establishing constituency relations between the words and semantic relations between the corresponding meanings.

Furthermore, language is used in communicative contexts and therefore language understanding needs to be appropriately grounded in the discourse and situational context. Here the father is coaxing the sick child to take her medicine and has used both gestures and actions to establish the cough syrup as their joint attention. This information is captured in Figure 1.1 on the right where the father (FAT) is the speaker, the child (CHI) is the addressee, and the cough syrup is the attentional-focus of the DISCOURSE_SEGMENT. The hearer's job is to link the sentence in with context and recognize that the speaker is asking specifically for her, the addressee, to be the eater of not just any medicine, but the specific cough syrup that they are jointly attending to. These links to context are represented in the figure using bold dashed arrows from the Child to CHI and from the eatee to Cough Syrup, etc. This kind of link to context is particularly critical for the hearer to understanding other sentences in which arguments are omitted.

A grammar learner's job is to find a systematic way of turning sequences of words such as *xi1xi1 chi1 yao4* into coherent interpretations such as the one just shown. Specifically, there are at least 4 pieces of linguistic knowledge necessary for this task:

1. The word *xi1xi1* is a label for a child with the name XiXi
2. The word *yao4* is a label for medicines
3. The word *chi1* is a label for the action of eating, which involves two participants, the eater (some human) and the eatee (some food), and the motor program of putting food in one's mouth and swallowing it
4. The sequence object-label – action-label – object-label means that the first object is performing the action to the second object.

By design, the model in this dissertation assumes knowledge of object and action labels such as (1) – (3) at the start and learns the language-specific ways to express relational meanings, i.e. argument structure constructions, such as those in (4). As pointed out by Givón (2001), languages routinely use a combination of intonation, word order, and morphology (in the form of verb agreement and nominal case marking) to mark grammatical relations (and by extension semantic relations). Without the computational support of a morphological analyzer[1] or a speech

---

[1] The support for inflectional morphology is not in place in ECG or the language understanding system at the time of this dissertation work, but ongoing efforts, in particular by Nathan Schneider, are being made in the group to extend ECG.

recognition system, this model is capable of handling only word order and free morphemes. Fortunately this does not affect the use of the model on Mandarin Chinese data, which only utilizes those two cues.

Two possible immediate objections to this simplifying assumption are to the apparent sequential nature of the learning and the rich verb semantics that the learner has access to prior to syntax. It is a fact that there is no such point in time during language development when word learning stops and grammar learning begins. In no way is subsequent or concurrent word learning precluded by the current model; the model merely begins at a point at which children have learned enough words to begin positing syntactic and semantic relations between them. The vocabulary size is kept constant to keep the model simple, but it is a straightforward manipulation to gradually expand the vocabulary of the model as learning progresses.

As for the second concern, for practical reasons[2] verb-specific schemas are used, but the learning algorithm itself does not depend on the particular shape of the schema hierarchy. Furthermore, there is still a lot more about grammar learning that is of interest despite assuming relatively precisely-defined action labels. These initial verbs, which are tied to embodied experiences of actions and could have been learned through a Bayesian learning process of the sort modeled by (Bailey, Feldman, Narayanan & Lakoff, 1997), make no claims about how they are used in conjunction with their arguments. To give a few concrete examples in Mandarin Chinese, the word *mo3* ('to apply') is associated with the motor program of APPLY, which involves three participants: the applier, the substance applied, and the surface that the substance is applied to. One can say *mo3 you2* ('apply lotion') just as well as *mo3 lian3* ('apply face'), and to express

---

[2] A handwritten, adult-like grammar was created to evaluate how well the language understanding system performs under near optimal circumstances. The same semantic types are used in the learner so that the comparison of the learned grammar to this baseline would be meaningful.

both the substance and the surface, one may use the object-marking coverb *ba3*, as in *ba3 you2 wang3 lian3 shang4 mo3* (‘CV_{obj} lotion CV_{dest} face LOC apply’). These sorts of argument structure constructions are intricate and are not an automatic consequence of *mo3* referring to the gestalt notion of transporting some substance from one place to the next — the word *cheng2* (‘to ladle’, roughly) seems to only allow the substance as the direct object. Knowledge about verb argument structures such as these is exactly the kind of grammatical knowledge pursued in this dissertation. Closely related to the issue of verb meaning is of course concept development throughout the period of language learning, which will most certainly impact the kind of linguistic distinctions that a child is able to make. We will return to the implications of both vocabulary and concept development for the learning model in the final chapter.

Acknowledging that comprehension and production both play important roles in language learning, this work focuses on the task of learning grammatical constructions through an iterative process of trying to better understand language in meaningful communicative contexts. The learner, as mentioned, starts out with an initial vocabulary of object labels and action labels. Since the construction grammar framework provides a parsimonious representation for knowledge of words and phrasal structures (i.e. they are all constructions), the initial vocabulary is given to the model in the form of a grammar that consists only of lexical constructions.

To satisfy a technical requirement of the best-fit constructional analyzer, the learner model has to assume a few grammatical types that can be the basic units of analysis. Conceptually, these are units separated by some word, phrase, or sentence boundaries so that the analyzer has natural stopping points in estimating its expectation of the next word. In the starting grammar these units are `MORPHEME`, `PHRASE`, and `CLAUSE`. All the words that the learner knows in the

with the basic model. The reduction in resolution score is apparent on the seen validation data but the performance on the unseen data is largely unaffected. Here is a possible explanation: The learner creates lexically-specific constructions to accommodate each piece of training data that the learner encounters. These lexically-specific constructions are instrumental in obtaining a good resolution score on the seen validation data (which is a subset of the training data that the learner just learned about) but are less useful for unseen data. Due to the situation-specific nature of the parent-child conversation, a number of these lexically-specific constructions do not get used again for a long time after they are learned and are thus purged by the decay operation. This causes the resolution score on the seen validation data of the "with decay" model to drop.

On the other hand, much of the unseen validation data requires generalizations that are made across the lexically-specific constructions. Since these generalizations are more recently created than the specific ones and are more widely applicable, they are less likley to be removed by the decay operation. As a result, the peformance on the unseen validation data is maintained.

Variation 2: lowering the statistic update discount factor

An attempt is made to understand how the probability mass given to newly learned constructions affect learning. In this second variation, the discount factor γ used in the statistics update is reduced while keeping the rest of the settings the same as variation 1. Recall from Chapter 6 that γ can be understood in terms of how confident the model is in its ability to learn the correct construction during each learning operation. The basic model and variation 1 both use $\gamma = 1.0$, which corresponds to an extremely confident learner. Here variation γ was reduced to 0.2 and the decay operation was adjusted correspondingly so that constructions modified over 50 learning episodes ago which have been used fewer than 0.6 times are purged. This was a necessary modification to prevent most compositions from quickly being purged.

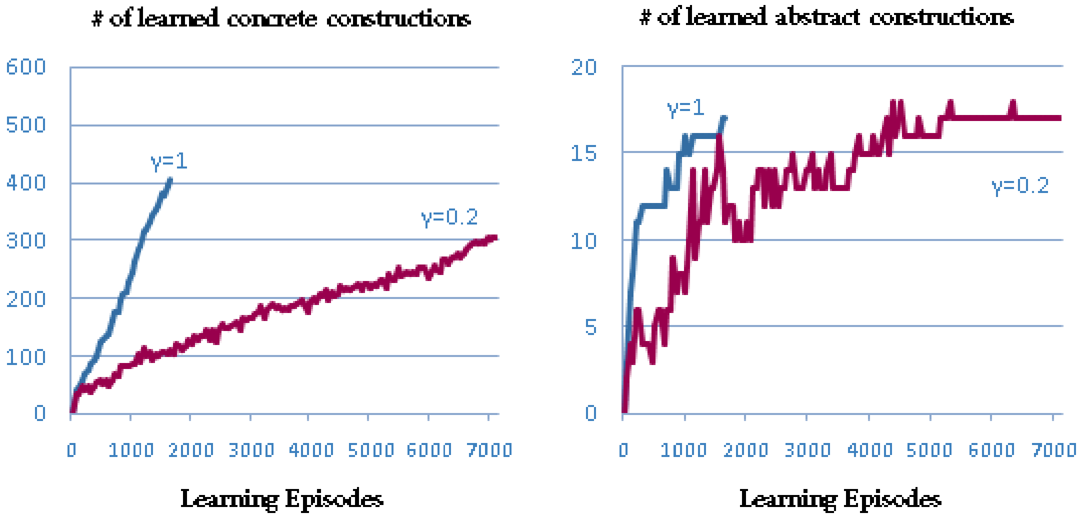

Figure 7.21 The size of the grammar grows with more fluctuations when γ = 0.2 than when γ = 1.0. The grammar seems to stabilize right at around 17 abstract categories though the number of concrete constructions continues to increase throughout.

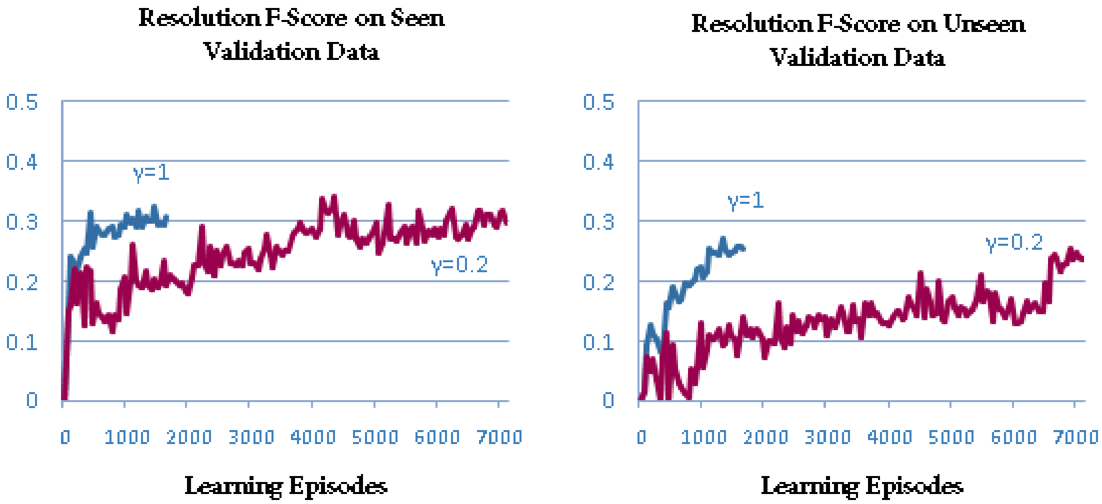

Figure 7.22 The resolution scores obtained on both datasets with γ = 0.2 are comparable to those with γ = 1.0 even though it takes the learner much longer to arrive at that level of performance.

The grammar size fluctuated more between learning episodes when γ = 0.2 compared to when γ = 1.0, which is reasonable since new generalizations and categories were less likely to be used with the lower discount factor and were more likely to be purged. It took the learner much longer to reach a similar level of performance on either the seen or unseen validation set, although it is worth pointing out that this model variation was able to get a better resolution score on the unseen data with fewer constructions than the basic model. Using 18 learned categories and 515 learned concrete constructions, the basic model obtained a resolution score of 0.208. With 17 learned categories and 313 learned concrete constructions, this variation had a score of 0.237.

The 17 constructional categories in the grammar learned by the "$\gamma = 0.2$" model, shown in the lefthand column of Figure 7.23 along with their semantic restrictions, are quite sensible. The semantic distinctions made by the categories resembled those made by the argument structure constructions in the handwritten adult grammar from Chapter 2. The "$\gamma = 1.0$" model also settled on 17 (different) constructional categories. They are shown in the righthand column of Figure 7.23 for comparison.

Three observations can be made about the constructional categories in the two learned grammars: (1) The two grammars have only 14 categories in common. The categories that are unique to each grammar are shaded in gray in the figure. Of the differing categories, those in the "$\gamma = 0.2$" grammar is semantically more general : `CAT9338 (INTRANSITIVE_ACTION)` as compared to `CAT1853 (SIT)`, and `CAT9624 (@Concrete_Entity)` as compared to `CAT1713` (`@Solid`). This may have to do with the fact that the "$\gamma = 0.2$" model was able to run for about 4 times as many iterations and therefore had more chances to form generalizations. (2) The categories in the two models are formed in different progressions even though the learning input is presented in the same order. (3) Two categories in the "$\gamma = 1.0$" grammar, `CAT442` and `CAT876`, remained subcases of another category, `CAT2496`. Recall that when one category A is created as a supertype of another category B during learning operations such as generalization or category expansion, the two categories are automatically merged unless the merge results in the breaking of semantic guarantees made by category B. The latter scenario is exactly what happened: a number

@ 0.2	@ 1.0
CAT038 INTRANSITIVE_STATE	CAT044 INTRANSITIVE_STATE
CAT300 TRANSLATIONAL_SELF_MOTION	CAT120 NEGATION
CAT2007 SOURCE_PATH_GOAL	CAT326 TRANSLATIONAL_SELF_MOTION
CAT2051 TRANSLATIONAL_FORCEFUL_MOTION	CAT393 TRANSLATIONAL_FORCEFUL_MOTION
CAT2240 FORCE_APPLICATION	CAT442 SUBCASE OF CAT2496 COMPLEX_TRANSITIVE_MOTOR_ACTION
CAT2367 TRANSITIVE_MOTOR_ACTION	CAT796 COMMUNICATION
CAT2679 INGESTION	CAT851 CAUSE_CHANGE
CAT3017 CAUSE_CHANGE	CAT876 SUBCASE OF CAT2496 FORCE_APPLICATION
CAT7673 NEGATION	CAT1713 @Solid
CAT8980 PERCEPTION	CAT1814 PERCEPTION
CAT9338 INTRANSITIVE_ACTION	CAT1853 SIT
CAT2933 SELF_MOTION	CAT2343 TWO_PARTICIPANT_STATE
CAT9469 TWO_PARTICIPANT_STATE	CAT2496 TRANSITIVE_MOTOR_ACTION
CAT9662 @Human	CAT2694 INGESTION
CAT9624 @Concrete_Entity	CAT3022 SELF_MOTION
CAT13409 UNCATEGORIZED_TRANSITIVE_ACTION	CAT3136 @Human
CAT16302 COMMUNICATION	CAT4077 UNCATEGORIZED_TRANSITIVE_ACTION

Figure 7.23 The 17 constructional categories in the grammar learned by the $\gamma = 0.2$ model are shown on the left, and the 17 categories in the grammar learned by the $\gamma = 1.0$ model are shown on the right. The categories that have no equivalent in the other grammar are shaded in gray.

of constructions were built around CAT876 and set up bindings with the force_supplier and force_recipient roles in its FORCE_APPLICATION schema. These two roles are no longer visible if the category is to be merged along with CAT442 into CAT2496, which has a meaning of TRANSITIVE_MOTOR_ACTION. These kinds of situations were avoided by the "$\gamma = 0.2$"model, whose slower path to generalization seemed to have allowed the grammar more wiggle room before settling.

Variation 3: perfect context-fitting

The central hypothesis in this learning model is that the ability of the learner to utilize contextually-obtained information is an enabling component of grammar learning. Though this hypothesis cannot be tested directly in this model (given that no learning can take place in this model without some way of determining the semantic relations between words), questions can still be asked about how important the accuracy of contextual inference is to the learner. Do initial learning mistakes based on imperfect intention reading hurt the learner in the long run?

The variation 1 model (i.e., basic + decay) was thus modified to take advantage of the gold standard annotation in the seen validation data. Specifically, the model used only the short dialogues as training data and the context-fitting process was tweaked to return only cotextual references that are consistent with the gold standard annotation. The learning outcome of the "perfect-knowledge" model was contrasted with that of model variation 1 re-run using only the short dialogues as training input.

To ground this comparison, the event core argument contextual fit scores of both models on the training data using a lexicon-only grammar were obtained. Recall that this score is a measure of how well the context-fitter performs. An evaluation experiment is run so that the learning model retreived up to 5 best analyses from the analyzer and re-ranked them using either the unmodified context-fitting process or the goldstandard-based context-fitting process. The top analysis (i.e. the one to be used for learning) is scored externally for their contextual fit and the respective scores for the two models were:

	core argument contextual fit		
	precision	recall	f-score
unmodified context-fitter	0.584	0.557	0.570
goldstandard context-fitter	0.991	0.905	0.946

Figure 7.24 The accuracy of the basic context fitter in the variation 1 model versus the goldstandard context fitter in the "perfect-knowledge" model. It can be very well expected that a lot of noise is introduced into the basic learning model. The learning outcomes of the two models are compared.

The f-score of the unmodified context fitter was just around 0.570, injecting noise into the learning process. A number of spurious constructions can be expected to be hypothesized by the learner based on the incorrect information. In contrast, the gold-standard context fitter had an f-score close to 0.95. (It was not at 1.0 because of lexical ambiguity: there is nothing that a perfect context fitter can do if the correct word sense is not in any of the analyses.)

We performed 4 iterations over the reduced training corpus (385 utterances, identical to the seen validation set) using both models. As expected, the resulting grammar in the variation 1 model is considerably larger than that in the "perfect-knowledge" model even as the number of abstract constructions (structured largely by the schema lattice) remained roughly equal. The "perfect-knowledge" model marginally outperformed the variation 1 model in the seen validation set but underperformed it in the unseen data. There could be two reasons why the variation 1 model with the faulty basic context fitter managed to do just as well as the one with perfect knowledge. The first is that there were enough "good" compositions created in the learning process that over time the good constructions resulted in useful generalizations and erroneous ones were unsed and purged. The second is that some of the spurious constructions were still consistent with Mandarin Chinese grammar even if they were inappropriate for the contextual situations in which they were learned. These spurious constructions nonetheless helped the learner make a first guess at the unseen data.

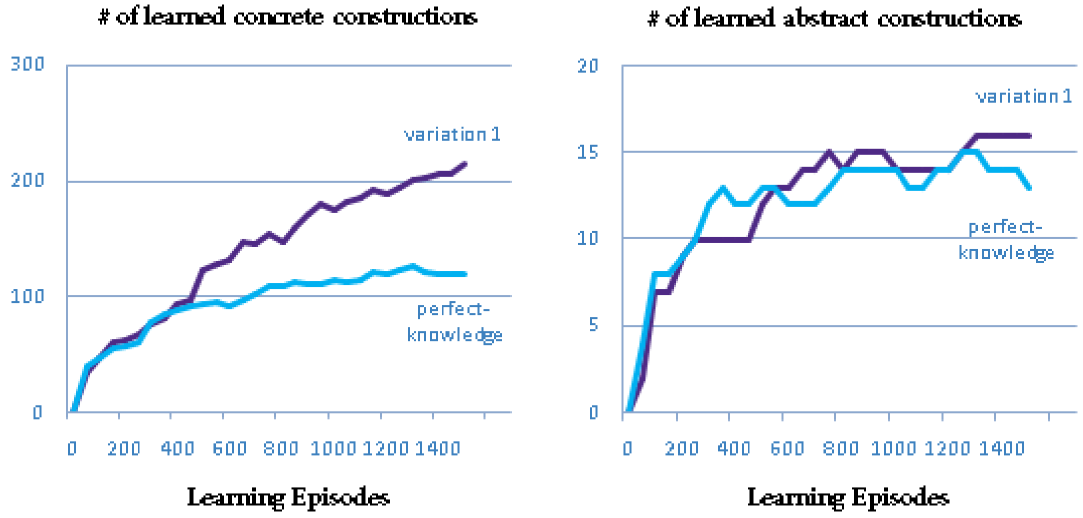

Figure 7.25 The number of learned concrete constructions after 4 learning iterations is significantly higher for the variation 1 model than the "perfect-knowledge" model, reflecting the amount of spurious constructions in the grammar.

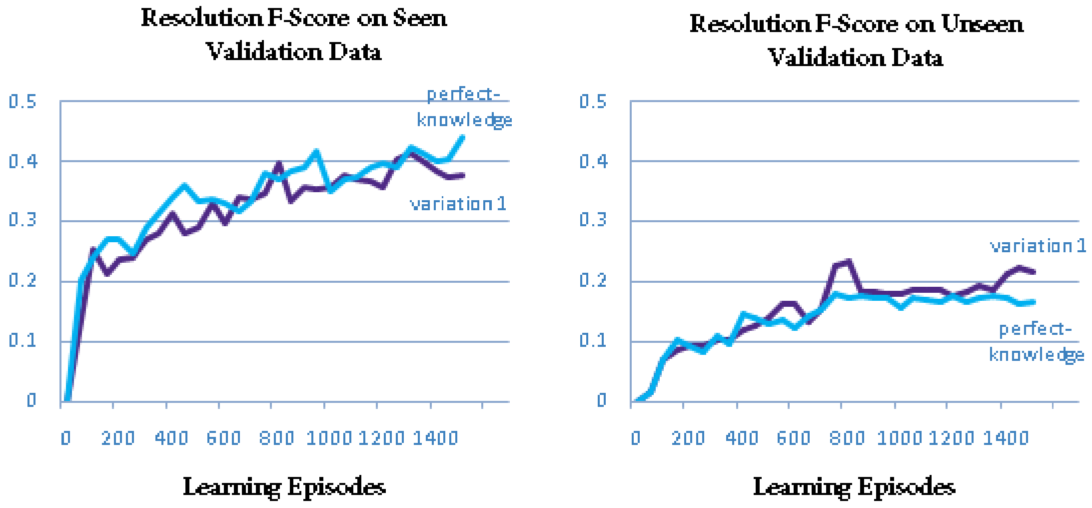

Figure 7.26 The resolution score on the seen and unseen validation data by the variation 1 vs "perfect-knowledge" model.

This result speaks to the robustness of the learning model, but two important questions about the learning model remain. Firstly, what is the contribution of each learning operation. Secondly and more poignantly, the learning results obtained here are far from that expected of a competent language user. How does a learner get from here to there? We attempt to address the first question in the next chapter using controlled experiments with a miniature artificial grammar. The second question is the holy grail of language development research, and this dissertation can only offer a discussion of the requirements and roadblocks to answering that question in the last chapter.

Chapter 8.

Artificial language learning experiments

The previous chapter describes two learning experiments using a subset of the Mandarin CHILDES corpus. While these experiments provide a strong demonstration of how the learner behaves on naturalistic data, the corpus is too big and noisy to systematically study the contribution of each learning operation. This chapter describes another set of experiments where we examine more closely the behavior of the learner model using a miniature Mandarin-like grammar and noiseless data.

In experiment 3, a simplistic SVO language with argument omission was used and both the combination of learning operations and the amount of training data were varied. The first manipulation, varying the combinations of learning operations used by the model, directly compares the contribution (or detraction, as the case may be) of different operations and examines their effect on the size and quality of the grammar eventually obtained. The second manipulation, varying the amount of training data, is performed with two objectives. The first goal is to understand how the availability of learning data relative to the complexity of the language affects learning outcome. The second goal is to better understand the contributions of the refinement operations; to push the envelope, so to speak, to see if these operations make more of a difference when learning input is scarce.

In a follow-up experiment, experiment 4, the miniature grammar was made more complicated by allowing object fronting and the learning results are compared against those in experiment 3. This is to draw attention to the problem both of function particles and more variable word order. This manipulation is a more direct test of the revision operation and is suggestive of other possible learning operations, which will be discussed in the final chapter.

8.1 Experiment 3: Mandarin-like artificial language learning experiment

Learning data

The miniature Mandarin Chinese-like language consists of 12 verbs, 20 nouns and no other function words (see Figure 8.1). Verbs fall in three semantic classes: intransitive states and actions, transitive states and actions and transfer. The noun meanings corresponding roughly to four groups: people, food, objects, and pictures. English words are used as the orthography to make the examples easily understood by non-Mandarin speakers; the intended corresponding Chinese words are provided in the figure for reference. No efforts were put into eliminating phonological cues to word classes or accounting for differences in frequencies of the words or concepts in real life as these are irrelevant to the current computational model. These words and their corresponding meaning are represented in ECG as constructions and schemas. Referent descriptors (RDs) are not used in the noun meanings to keep the grammar simple. A few representative schemas and constructions are shown in Figure 8.2.

The training and validation data in this language were obtained by first generating all the semantically plausible situational contexts using the available processes and entities, creating a total of 860 unique scenes. Example semantic restrictions include animacy requirements for processes such as SLEEP and FIND, disallowing reflexives in the transitive scenes, and an object type constraint (excluding pictures) on throwable items. Each scene is paired with an utterance in the appropriate intransitive / transitive / ditransitive frame as described by these three rules (which have the same semantics as English or Mandarin):

	verbs			nouns	
orth	**meaning**	**Chinese**	**orth**	**meaning**	**Chinese**
pretty	Pretty	*mei3*	I	@Xixi	*wo3*
			You	@Human	*ni3*
sleep	Sleep	*shui4*	Aunt	@Aunt	*a1yi2*
come	Come	*lai2*	haoyu	@Haoyu	*hao2yu3*
fall	Fall	*dao3*	dad	@Father	*ba1*
like	Like	*xi3huan1*	rice	@Rice	*fan4*
			apple	@Apple	*ping2guo3*
eat	Eat	*chi4*	fish	@Fish	*yu3*
disturb	Disturb	*nong4*	meat	@Meat	*rou4*
throw	Throw	*reng1*	veggies	@Vegetables	*cai4*
take	Take	*na2*			
find	Find	*zhao3*			
watch	Watch	*kan4*	ball	@Ball	*qiu2*
			book	@Book	*shu1*
			pen	@Pen	*bi3*
give	Give	*gei3*	car	@Car	*che1*
			stool	@Stepstool	*ban3deng4*
			bear	@Bear	*xiong2*
			horse	@Horse	*ma3*
			giraffe	@Giraffe	*lu4*
			stars	@Stars	*xing1xing1*
			moon	@Moon	*yue4liang4*

Figure 8.1 The miniature language consists of 12 verbs and 20 nouns. Their meanings are represented as schemas and ontology types, respectively. The verbs on the left fall in three semantic groups: intransitive states/actions, transitive states/actions, and transfer. The finer distinctions between types of processes are shown in the process lattice in the next figure. The nouns on the right fall in four semantic groups: human nouns, food names, object names, depicted objects.

S -> N V_{intran}

S -> N V_{tran} N

S -> N V_{ditran} N N

Arguments in each utterance are omitted with the following probabilities: the pre-verbal noun at 0.7, the first post-verbal noun at 0.4, and the second post-verbal noun (in ditransitives) at

0.6. This produces 860 utterances with distinct situational meanings; utterances may have the same form due to omission. The miniature corpus is annotated with events, speech acts, and gold standard the same way as the CHILDES corpus used in the experiments in Chapter 7. A sample of the generated sentences is shown in Figure 8.3.

```
Schema   PROCESS
    roles
        protagonist: @Entity
```

```
Schema   STATE
    subcase   of  Process
```

```
Schema   INTRANSITIVE_PROCESS
    subcase   of  Process
```

```
Schema   INTRANSITIVE_STATE
    subcase   of  STATE,
    INTRANSITIVE_PROCESS
```

```
Schema   PRETTY
    subcase   of  INTRANSITIVE_STATE
```

```
construction     PRETTY
    subcase   of  MORPHEME
    form
        constraints
            self.f.orth <-- "pretty"
    meaning   : Pretty
```

```
Schema   ACTION
    subcase of PROCESS
```

```
Schema   TWO_PARTICIPANT_PROCESS
    subcase   of  PROCESS
    roles
        protagonist2: @Entity
```

```
Schema   TRANSITIVE_ACTION
    subcase   of  ACTION,
    TWO_PARTICIPANT_PROCESS
```

```
Schema   TAKE
    subcase   of  TRANSITIVE_ACTION
```

```
construction     TAKE
    subcase   of  MORPHEME
    form
        constraints
            self.f.orth <-- "take"
    meaning   : TAKE
```

Figure 8.2 A partial view of the process hierarchy and two words, ***pretty*** and ***take***, in the miniature language.

A sample of the sentences in the miniature language

ball pretty
sleep
like rice
I like fish

Haoyu disturb horse
throw
Aunt throw
take car

find fish
give car
I give
give

Figure 8.3 Example of sentences in the miniature language. Each of the sentence here has a unique associated situational context. An utterance such as *take car* may be observed twice in the data but the takers in the two scenes are different.

Training procedure

To verify the integrity of the data , a randomly subset of the resulting sentences were tested on the analyzer using a handwritten grammar containing only the three subject verb phrase constructions. These sentences were correctly analyzed as long as (1) each construction poses semantic limitations on the verbs and (2) arguments were allowed to be omitted at the specified rate[38, 39]. This establishes the ceiling of the resolution score measure (see Section 7.2) at 1.0 and the floor of the average number of roots per analysis measure (see Section 7.2) at 1.0.

It is important to keep in mind that even though this miniature grammar is created using 3 basic syntactic frames, in principle even without constructions that allow omission, 14 clausal constructions are sufficient to analyze the data (2 for the two omission patterns in the intransitive frame, 4 for the transitive frame and 8 for the ditransitive frame). What argument omission does is to allow for compactness and parsimony in the grammar.

A validation set was created by randomly selecting 20% of the 860 valid sentences. The remaining 688 sentences were available as training data in a set of experimental runs. The first manipulation was the size of the training corpus: percentages (5%, 25%, 50%, and 100%) of the 688 sentences were randomly selected as the training set. The second manipulation was the combination of learning operations used:

I. composition only
II. composition + generalization
III. composition + generalization + decay

---

[38] Additional settings: the analyzer has to be forced to return single-rooted parses and individual morphemes must be disallowed from being the root of an analysis. Otherwise the analyzer in some occasions choose instead multi-rooted parses over omitting multiple arguments, or choose a single verb as the final analysis. These problems are consistent with those results achieved in [Bryant, 2008].

[39] Only one sentence, *give book*, generated an incorrect analysis. Without semantic constraints on the schema roles of give, the book was analyzed as the recipient rather than the theme.

IV. composition + generalization + category expansion + decay

V. composition + generalization + revision + decay

VI. composition + generalization + omission + decay

VII. composition + generalization + category expansion + revision + omission + decay

Variation I is the absolute minimum for the learner. Variation II has just the basic composition and generalization mechanisms. Variation III introduces the decay mechanism, which we found from Section 7.3 to be useful in keeping the grammar size under control. This is what we will be calling the "no refinement" model. Variations IV through VII are the "refinement" models with different combinations of refinement operations. Variation VII is the same as the basic model introduced in the last chapter with all operations enabled. Category merges triggered by generalization and category expansion are allowed in all variations.

The total of 28 model variations were each run twice with two different randomly selected subset of the training corpus with a maximum of 6 iterations. The results across the two runs are averaged in the quantitative results reported here. The statistic update discount factor γ was set to 0.5, noncompositional meaning or maximally-connected compositions were disabled, and a uniform semantic model was used. The learning model obtained up to 5 best analyses from the best-fit analyzer using a multi-root penalty of -20 (in log probability scale). Since the primary interest here is to examine the contribution of each learning operation, the gold standard context fitter was used to eliminate noise just as in the "perfect knowledge" variation of the basic model (variation 3) from the previous chapter.

Quantitative results

We will begin with the results of using all available training data, that is, a training set size of 688 in contrast to a validation set size of 172. The resolution scores of the seven learning operation combinations are reported in Figure 8.4. As expected, the variation I model with only the composition operation does not generalize very well to the validation set. The best core argument resolution results using this dataset is obtained by using composition, generalization, and decay in conjunction with constituent omission, although it does take the model a few iterations longer to reach that level of performance. In terms of both the resolution score and the average number of root per analysis measures, however, all the model variations that allow generalization perform roughly equally well.

	I	II	III	IV	V	**VI**	VII
iter	**Comp**	Comp+Gen	CG+Dec	CGD +Exp	CGD+Rev	**CGD +Oms**	CGD+ Exp+Rev+Oms
0	**0.000**	0.000	0.000	0.000	0.000	**0.000**	0.000
1	**0.385**	0.836	0.836	0.826	0.849	**0.766**	0.762
2	**0.388**	0.854	0.851	0.836	0.843	**0.828**	0.785
3	**0.398**	0.855	0.847	0.829	0.848	**0.890**	0.791
4	**0.395**	0.851	0.849	0.825	0.845	**0.918**	0.806
5	**0.392**	0.845	0.851	0.836	0.855	**0.914**	0.807
6	**0.395**	0.851	0.856	0.839	0.851*	**0.907**	0.807*

Figure 8.4 The resolution score obtained over the course of 6 iterations using 100% of the training data. The best scores were achieved by variation VI and the worst by variation I, both highlighted in bold. The seven different combinations are as described on page 179, abbreviated here as: Comp = Composition, Comp Gen (CG) = Composition + Generalization, CG Dec (CGD) = Composition + Generalization + Decay, Exp = Category Expansion, Rev = Construction Revision, Oms = Constituent Omission.

*** run ended early[40] and the score from the last learning episode is reported instead. The star denotes shortened runs in subsequent tables.**

---

[40] An analysis may be returned by the analyzer even when some of the RDs have no existing compatible referents — this is by design because language may introduce new referents. Consequently, it is sometimes possible that none of the top analyses returned by the analyzer are compatible with context. These incompatibilities are caught by the gold-standard fitter (but not the basic fitter) and the learning episode is skipped as a result.

	I	II	III	**IV**	V	VI	VII
iter	**Comp**	Comp+Gen	CG+Dec	**CGD +Exp**	CGD+Rev	CGD +Oms	CGD+ Exp+Rev+Oms
0	**2.01**	2.01	2.01	**2.01**	2.01	2.01	2.01
1	**1.34**	1.09	1.13	**1.03**	1.18	1.10	1.06
2	**1.32**	1.09	1.11	**1.03**	1.10	1.10	1.06
3	**1.32**	1.09	1.11	**1.03**	1.10	1.10	1.06
4	**1.32**	1.09	1.11	**1.03**	1.09	1.10	1.06
5	**1.32**	1.09	1.11	**1.03**	1.09	1.10	1.06
6	**1.32**	1.09	1.11	**1.03**	1.09*	1.10	1.06*

Figure 8.5 The average number of roots per analysis over the course of 6 iterations using 100% of the training data. The most cohesive analyses were achieved by variation IV and the least by variation I, highlighted in bold.

Looking at the number of concrete constructions learned across the 7 different combinations of learning operation using all the available training data, it is immediately apparent that variation I led to an order of magnitude more constructions than any other combinations that include generalization. However, as previously discussed, only 3 constructions are strictly necessary to analyze the data (14 if omission is not allowed). All model variations learned many more constructions than are strictly necessary to analyze the data, reflecting (1) idiosyncrasies in the data (including semantic restrictions on the core arguments of processes), (2) a conservative learning approach that generalizes only as much as the data warrants, and (3) competition between the specific and general constructions that led to the preservation of a large number of specific constructions.

No constructional categories are created by the composition-only model, and as expected, category expansion led to bigger and therefore fewer categories. The seven categories learned by the composition + generalization model are (semantically): TRANSITIVE_ACTION, INTRANSITIVE_ACTION, @Object, @Object, @Object, @Inanimate, @Inanimate. These categories have overlapping members but their constructional contexts are distributed differently. On the other hand, with category expansion, the learned four constructional categories are (semantically): `TRANSITIVE_ACTION`, `INTRANSITIVE_ACTION`, `@Object`, and `@Inanimate`. In an alternate run with the same setting but a differently chosen training set, the category of inanimate nouns merged with the category of object nouns.

	I	II	**III**	IV	V	VI	VII
iter	**Comp**	Comp+Gen	**CG+Dec**	CGD +Exp	CGD+Rev	CGD +Oms	CGD+ Exp+Rev+Oms
0	**0**	0	**0**	0	0	0	0
1	**558**	164	**61**	76.5	64.5	68	71
2	**607**	163	**56.5**	76.5	63.5	75	89.5
3	**607**	163	**56.5**	76.5	63.5	77	85
4	**607**	163	**56.5**	76.5	63.5	76.5	83
5	**607**	163	**56.5**	76.5	63.5	77.5	84
6	**607**	163	**56.5**	76.5	63.5	77.5*	85.5*

Figure 8.6 The number of learned concrete constructions after 6 iterations using 100% of the training data. The largest grammar results from variation I and the smallest from variation III, highlighted in bold.

	I	**II**	III	IV	V	VI	VII
iter	**Comp**	**Comp+Gen**	CG+Dec	CGD +Exp	CGD+Rev	CGD +Oms	CGD+ Exp+Rev+Oms
0	**0**	**0**	0	0	0	0	0
1	**0**	**7**	5.5	3.5	5	4.5	3.5
2	**0**	**7**	5	3.5	5	4.5	3.5
3	**0**	**7**	5	3.5	5	4.5	3.5
4	**0**	**7**	5	3.5	5	4.5	3.5
5	**0**	**7**	5	3.5	5	4.5	3.5
6	**0**	**7**	5	3.5	5	4.5*	3.5*

Figure 8.7 The number of learned constructional categories after 6 iterations using 100% of the training data. Variation I results in no constructional categories, and the most number of categories are learned by variation II, highlighted in bold.

The per-iteration results obscure the more interesting changes in grammar size that happened within the first two learning iterations, which are plotted in Figure 8.8 and Figure 8.9. For the composition-only model, the number of learned concrete constructions increased linearly with the number of utterances encountered in the first iteration (slope = 1.05), and increased linearly but at a much slower pace during the second iteration (slope = 0.07). This reflects the fact that constructions learned in the first iteration are used to analyze utterances in the second iteration such that additional compositions can be performed.

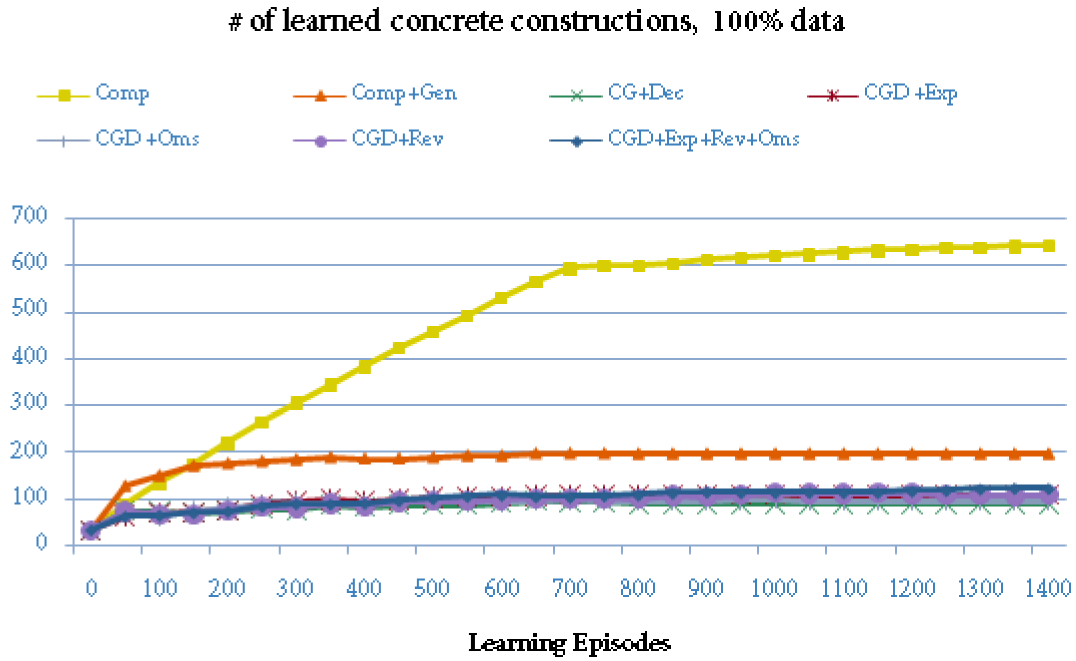

Figure 8.8 The number of learned concrete constructions as a function of the learning episodes.

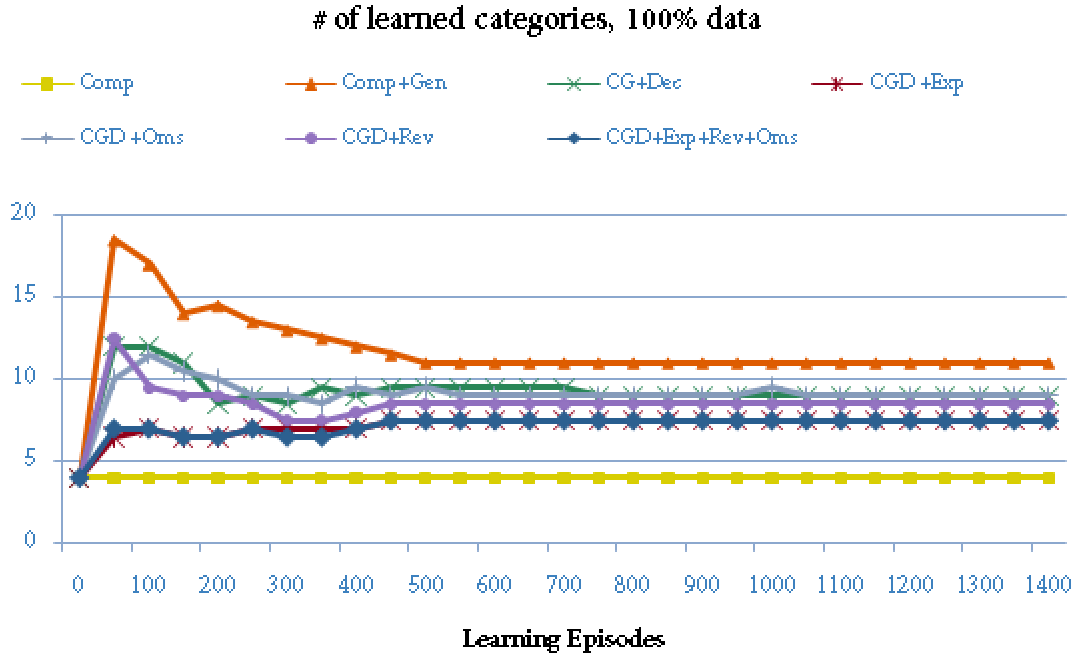

Figure 8.9 The number of learned categories as a function of learning episodes.

As for the number of constructional categories, model variations II, III, V, and VI (Comp+Gen, CG+Dec, CGD+Rev and CGD +Oms) all overshot before settling at the their final set of categories[41]. From a modeling standpoint, it is worth noting that even though all of the variations except for composition-only perform similarly well, the variations with the refinement operations were able to do so at less than half the grammar size of the Comp+Gen variation. This is an important consideration in light of the long-term memory demands that the grammar storage may pose on the learner.

The results reported so far are from using 100% of the 688 examples in the training data. The results from the second, training set size manipulation are presented here. The same seven variations of the model were trained on 50% (344), 25% (172) and 5% (34) of the available learning corpus and tested on a validation set of size 172. By both the core argument resolution measure (Figure 8.10) and the average number of roots per analysis measure (Figure 8.11), the best-performing variations in the 100% and the 50% conditions are largely comparable even as there are slight drop-offs in the other variations. However, the degradation in resolution score is much more noticeable between the 50% and the 25% conditions, and between the 25% and the 5% conditions.

% data	I Comp	II Comp+Gen	III CG+Dec	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	VII CGD+ Exp+Rev+Oms
100%	0.395	0.847	0.846	0.837	**0.908**	0.848	0.807
50%	0.344	0.805	0.847	0.795	**0.905**	0.814	0.751
25%	0.313	0.764	0.762	**0.787**	0.705	0.783	0.787
5%	0.194	0.350	0.390	0.565	0.344	0.466	**0.570**

Figure 8.10 The resolution score obtained at the end of 6 iterations for varying amounts of training data. The best performing model in each training set size condition is highlighted in bold.

---

[41] Even though decay is not enabled in the Comp+Gen variation, category merges triggered by constructional generalization are allowed to happen, and in the process ridding the grammar of excess categories.

% data	I Comp	II Comp+Gen	III CG+Dec	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	VII CGD+ Exp+Rev+Oms
100%	1.32	1.09	1.11	**1.03**	1.09	1.10	1.06
50%	1.46	1.14	1.14	1.05	1.19	1.13	**1.05**
25%	1.56	1.20	1.19	1.07	1.16	1.19	**1.04**
5%	1.81	1.66	1.70	1.31	1.69	1.65	**1.24**

Figure 8.11 The average number of roots per analysis obtained at the end of 6 iterations for varying amounts of training data. The best performing model in each training set size condition is highlighted in bold.

The basic finding here is not news. Having large corpora helps. The somewhat surprising result is that the 5% model did as well as it did (resolution score = 0.570) using a combination of composition, generalization, revision, omission and decay. Figure 8.12 breaks down the difference in resolution scores between all the "refinement" models IV-VII and the "no refinement" baseline of variation III. Whereas the refinement operations did not help or even hurt when there were large amounts of data (possibly due to overfitting), they were generally helpful when the data was very sparse. Furthermore, the learning in variation VII (all learning operations) using 5% of the training data resulted in 45 learned concrete constructions and 3 learned abstract constructions, as compared to 607 learned concrete constructions and 0 learned abstract constructions in variation I (composition only) using 100% of the data. This is impressive considering that the 5% variation VII model also did better than the 100% variation I model (0.570 compared to 0.395).

The average improvement of the refinement models is also greatest when there is the least amount of training data. Before we can extrapolate from these results and make conclusions about the difficulties in learning from real Mandarin Chinese data, however, we would like to look at how these operations scale with linguistic complexity of the language. For this we turn to another set of experiments using a modified and more complex miniature grammar.

% data	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	VII CGD+ Exp+Rev+Oms	Avg
100%	-0.009	0.002	0.063	-0.038	0.005
50%	-0.052	-0.034	0.058	**-0.096**	-0.031
25%	0.025	0.021	-0.057	0.025	0.004
5%	0.175	0.076	-0.045	**0.181**	0.097

Figure 8.12 The difference in resolution score of the "refinement" models from the "no refinement" baseline. The average improvements of the refinment models are also shown. The most and least improvement from the refinement models, as well as the most average improvement, are highlighted in bold.

8.2 Experiment 4: Mandarin-like artificial language with object fronting

Learning data

To start getting at what effect linguistic complexity has on the learning model, here we make a slight modification to the miniature grammar introduced in Section 8.1. In this version a new particle *ba3* is introduced as an object marker (just as in the Mandarin Chinese). Using this coverb marker an object can be moved to a preverbal position. In the transitive sentences, the fronted object is semantically the patient. In the ditransitive sentences, the fronted object is semantically the theme. As a result, disregarding argument omission, there can be two ways to construct a transitive sentence and two ways to construct a ditransitive sentence in this modified miniature language (with the corresponding semantic arguments in parenthesis):

S -> N V_{intran} (agent)

S -> N V_{tran} N (agent patient) S -> N ba3 N V_{tran} (agent patient)

S -> N V_{ditran} N N (giver recipient theme) S -> N ba3 N V_{ditran} N (giver theme recipient)

With the exception of the additional lexical construction BA3 (with empty meaning), the same vocabulary and semantic schemas from the original miniature language are used here and the same 860 situational contexts are reused here. The arguments in this miniature language corpus are generated using a slightly different procedure. A sentence with all arguments present is generated for each situational context. For the agent/giver constituent, $P(\text{expressed}) = 0.3$ and $P(\text{local} \mid \text{expressed}) = 1.0$. For the recipient constituent, $P(\text{expressed}) = 0.4$ and $P(\text{local} \mid \text{expressed}) = 1.0$. These are unchanged from the last miniature grammar.

The patient in the transitive frame and the theme in the ditransitive frame are the ones which can be fronted, and the data generation is a bit tricky. For the patient in the transitive frame, $P(\text{expressed}) = 0.4$ and $P(\text{local} \mid \text{expressed}) = 0.65$. For the theme in the ditransitive frame, fronting of the theme is only allowed when the recipient is also expressed, consistent with conventions in Mandarin Chinese[42]. This dependency of one constituent's locality on another constituent is not well captured by the mathematical model in the current analyzer. Nonetheless, the theme in the ditransitive scene is generated with $P(\text{expressedtheme}) = 0.6$, and if the fronting conditions are met (i.e. the recipient is also expressed), $P(\text{localtheme} \mid \text{expressedtheme}) = 0.35$. A few examples of the generated sentences that have object fronting are shown in Figure 8.13. The rest of the generated sentences look just like those shown in Figure 8.3.

A sample of sentences with fronted objects in the modified miniature language

ba3 Aunt like
ba3 Dad disturb
HaoYu ba3 I disturb
ba3 car throw
ba3 veggies find
you ba3 stool find
ba3 bear give you
Haoyu ba3 rice give I
Dad ba3 bear give Haoyu

Figure 8.13 Example of sentences in the modified miniature language.

---

[42] Certainly, in Mandarin Chinese there are other semantic restrictions and implications of fronting. One has to do with a notion of affectedness / disposal associated with the direct object (Li & Thompson, 1981). Another is the information structure of the sentence. Neither are taken into account in this miniature grammar to keep things simple.

Training procedure

The training procedure for this experiment is exactly the same as in experiment 3. The same two learning operation combination and training set size manipulations were used.

Qualitative results

Since revision is enabled only in variation V and VII, those are the only two sets of results where learned constructions with a BA3 constituent are expected. This was indeed the case. We focus on these constructions since the fronted object sentences make up the primary difference between this experiment and experiment 3. Figure 8.14 shows a sample of the ba3-using constructions learned by the variation V model using the same shorthand as in Chapter 7.

Resulting Construction	**Meaning Gloss + Contextual Restriction**
BA3-CAT1262-CAT154-c1066 CAT1262: inanimate nouns (excluding pictures) CAT154: transitive action verbs	<Human> - TRANSITIVE_ACTION - Inanimate
AUNT-BA3-CAT2149-LIKE-c2151 CAT2149: object nouns	Aunt - LIKE - Object
BA3-CAT1607-GIVE-CAT2149-c2585 CAT1607: inanimate nouns CAT2149: object nouns	<Human> - GIVE - Human - Inanimate
BA3-DAD-CAT154 CAT154: transitive action verbs	<Human> - DISTURB - Dad

Figure 8.14 Example s of constructions with BA3 learned by model variation V.

Once both revision and omission are thrown into the mix in the variation VII model, the constructions learned are unexpected at best and erroneous at worst. Figure 8.15 shows two of these constructions from each run of the model. Often the BA3 particle is learned to be optional while the direct object it marks is learned to be omissible. This happens an utterance with just the main verb is contrasted with a specific construction with a fronted object, and the learner marks the non-core BA3 optional and the core object omissible.

Resulting Construction	Meaning Gloss + Contextual Restriction
[BA3]-[CAT087]-LIKE-c766 CAT087: object nouns	<Human> - LIKE - Object
[BA3]-[CAT087]-CAT015-c766 CAT087: object nouns CAT015: transitive action verbs	<Human> - TRANSITIVE_ACTION - Object

Resulting Construction	Meaning Gloss + Contextual Restriction
[HAOYU]-[BA3]-[CAT020]-CAT063-c943 CAT020: object nouns CAT063: transitive action verbs	Haoyu - TRANSITIVE_ACTION - Object
[CAT020]-[BA3]-[CAT020]-CAT063-c976 CAT020: object nouns CAT063: transitive action verbs	Object - TRANSITIVE_ACTION - Object

Figure 8.15 Examples of constructions with BA3 learned by the two runs using model variation VII.

Quantitative results

The quantitative results confirm the idea that this grammar is more difficult for the model to learn than the grammar in Experiment 3. Figure 8.16 and Figure 8.17 show the performance of the models over the course of 6 training iterations on the modified grammar in contrast with the final results obtained by the same models in Experiment 3. The current results are worse across the board. The difference between scores obtained by the "no refinement" variation II and the scores obtained by the "refinement" variations IV through VII confirms the suspicion that a number of unwarranted generalizations are made with respect to omissible and optional arguments (as demonstrated in the qualitative results). This is also reflected in the reduced average improvements of the "refinement" models over the "no refinement" baselines compared to Experiment 3, as shown in Figure 8.18.

These results underscore how intuitive learning principles can sometimes produce unexpected or even incorrect results when used on a large scale. A separate set of data analyses was conducted to examine the likely causes. The 4 learned grammars from the two runs each of variations VI (CGD + Oms) and VII (CGD + Exp + Rev + Oms) were used to analyze all 860 sentences in this miniature grammar corpus. To gauge the amount of ambiguity in the grammars, the analyzer was asked to return a maximum of 15 analyses for each sentence in the corpus. If there is little ambiguity in the grammar, the average number of returned analysis for each sentence is expected to be close to 3 or 4 (1 analysis for the unambiguous single-rooted analysis and several much worse multi-rooted analyses). The average number of returned analyses per utterance using the varation VI grammars was 12.75, whereas the average number of returned analyses per utterance using the varation VII grammars was 13.00. This suggests that the combination of additional learning operations may have created constructions that are individually reasonable, but as a set added so much ambiguity in the grammar that they undermine the analyzer's ability to pick out the correct analysis.

	I	II	III	IV	V	**VI**	VII
iter	**Comp**	Comp+Gen	CG+Dec	CGD +Exp	CGD+Rev	**CGD +Oms**	CGD+ Exp+Rev+Oms
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.335	0.809	0.786	0.770	0.794	0.694	0.738
2	0.333	0.821	0.795	0.780	0.807	0.765	0.753
3	0.330	0.823	0.787	0.787	0.813	0.833	0.758
4	0.330	0.819	0.790	0.785	0.797	0.858	0.749
5	0.327	0.818	0.787	0.787	0.802	0.810	0.753
6	0.330	0.821	0.795	0.781	0.790	**0.867***	0.747*
Exp 3	0.395	0.847	0.846	0.837	**0.908**	0.848	0.807

Figure 8.16 The resolution score obtained over the course of 6 iterations using 100% of the training data. The best scores were achieved by variation VI and the worst by variation I, both highlighted in bold. These are contrasted by the results obtained on the simpler grammar in Experiment 3.

	I	II	III	IV	V	VI	**VII**
iter	**Comp**	Comp+Gen	CG+Dec	CGD +Exp	CGD+Rev	CGD +Oms	**CGD+ Exp+Rev+Oms**
0	**2.18**	2.18	2.18	2.18	2.18	2.18	**2.18**
1	**1.63**	1.31	1.32	1.22	1.22	1.45	**1.13**
2	**1.63**	1.30	1.29	1.20	1.20	1.42	**1.10**
3	**1.63**	1.30	1.29	1.20	1.19	1.37	**1.10**
4	**1.63**	1.30	1.29	1.20	1.19	1.36	**1.10**
5	**1.63**	1.30	1.29	1.20	1.19	1.36	**1.10**
6	**1.63**	1.30	1.29	1.20	1.19	1.378*	**1.14***
Exp 3	1.32	1.09	1.11	**1.03**	1.09	1.10	1.06

Figure 8.17 The average number of roots per analysis over the course of 6 iterations using 100% of the training data. The most cohesive analyses were achieved by variation VII and the least by variation I, highlighted in bold. These are contrasted by the results obtained on the simpler grammar in Experiment 3.

	IV	V	VI	VII	Avg	Exp 3 Avg
% data	CGD +Exp	CGD+Rev	CGD +Oms	CGD+ Exp+Rev+Oms		
100%	0.000	0.018	0.079	-0.041	0.014	0.005
50%	-0.002	0.004	-0.082	0.050	-0.007	-0.031
25%	-0.042	0.008	0.029	-0.026	-0.008	0.004
5%	0.161	0.035	-0.032	0.099	0.066	0.097

Figure 8.18 The difference in resolution score of the “refinement” models from the “no refinement” baseline in the more complex grammar. The average improvements of the refinment models are also shown. The most and least improvement from the refinement models, as well as the most average improvement, are highlighted in bold. These are contrasted with results obtained on the simpler grammar in Experiment 3.

The rest of the behavioral patterns of the model discussed in Section 8.1 hold in this experiment and the discussion of those results are therefore omitted from this dissertation. These two experiments represent a first step in understanding the behavior of a complex learning system. The final chapter discusses more approaches of combining the power of using naturalistic data and miniature languages in this computational framework.

Chapter 9.

Discussion and Future Directions

At the beginning of this dissertation, the following question was posed: if natural languages are too complex to be learned by blind associations, what is the nature of the innate learning biases that human learners are endowed with such that they almost always learn their native languages successfully?

This dissertation explored structural learning biases in the form of

€ a child learner's understanding that forms (words and phrases) have referential meaning and her desire to make sense of the utterances,

€ situational information that informs the possible interpretations of utterances, and

€ embodied semantic knowledge that establishes semantic coherence in learned constructions and guides the generalization of constructions

All of these structural biases are extremely effective in reducing the hypothesis space for new constructions, which are created in the learning model through a combination of utterance-dependent and utterance-independent learning operations. The utterance-dependent learning operations, which directly utilize the output of the best-fit constructional analyzer, include the basic operation *composition* and the refinement operations *construction revision* and *constituent omission*. The utterance-independent learning operations, which manipulate existing constructions in the grammar, include the basic *generalization* operation and the refinement operations *category merge*, *category expansion*, and *decay*. The result is a comprehension-driven learning framework that simultaneously learns both grammatical structures and statistical parameters on these grammatical structures.

To evaluate a cognitive modeling framework for grammar learning such as this, four criteria must be met. First, the model must display the same general learning tendencies as a child learner. Second, the model must be able to learn correctly under a variety of circumstances. Third, the model must have clear assumptions and systematic model parameters. Forth, the model should be well-motivated in implementation such that it is extendible beyond its initial modeling goals.

The rest of this chapter will address each of these criteria in turn. Section 9.1 answers to the first two criteria by looking across the results from the four learning experiments with naturalistic data as well as artificial languages. Section 9.2 takes up the issue of modeling assumptions and model parameters by looking at constructional generalization as a case study. Section 9.3 looks at two additional kinds of constructions that have not been the primary focus of the model — constructions with non-compositional meaning and function morphemes — and discusses how the model can be reasonably extended to model the learning thereof.

Finally, as any good thesis should, Section 9.4 offers some wild speculations about what this dissertation might have to do with a host of related issues, such as word learning, concept learning, morphosyntactic development, and the general problem of using situational context for language understanding and learning.

9.1 General discussion of the natural language and artificial language experiments

Learning sequences performed by the model discussed in Section 7.3 give assurance that the model is making reasonable learning choices given a corpus of real parent-child interaction. Among the learned constructions are:

- a proto NP-VP construction with a meaning of an INTRANSITIVE_STATE, where the NP-like constituent is a category of words that refer to @Solids and has a probability of 0.46 of being expressed, and the VP-like constituent is a category of INTRANSITIVE_STATE words.
- a proto VP-NP construction with a meaning of FORCE_APPLICATION, where the VP-like constituent is a category of FORCE_APPLICATION words and the NP-like constituent is the same category of words that refer to @Solids as the above.
- another proto VP-NP construction denoting a TWO_PARTICIPANT_STATE, where the VP-like constituent is a category of stative relation verbs and the NP-like constituent is a category of human-referring words and has a probability of 0.37 of being expressed. There are also other more lexically-specific constructions with a meaning of TWO_PARTICIPANT_STATE that have 3 constituents.
- a proto NEG-VP construction with a central scene of INGESTION, where the negation word is optionally expressed with a probability of 0.27 and the VP-like constituent is a category of INGESTION words.

Additionally, the model also learned subcategorization preferences for each of these constructions. For example, in the intransitive NP-VP-like construction, the word *huai4* (broken) was used 63.6% of the time (14 out of 22). The less frequently used verbs are *hao3* (good), *xiao3* (small), and *hao3wanr2* (amusing). This is consistent with the work of Wonnacott et al.(2008) which suggests that adult learners learn probabilistic subcategorization constraints.

These relatively general constructions along with more specific ones give the learner tremendous leverage in understanding new utterances, drawing out not only their phrasal structures but also their semantic bindings. In terms of the macro behavior, the comprehension-driven grammar learning model does what it sets out to do: to understand each piece of learning input as best it can based on its current grammar and the situational context, and compose a new construction if there are form-meaning mappings not captured by any constructions in the current grammar. As such we expect the number of composition operations to decrease over time as the learner's grammar gains coverage, and this is exactly what we see in the experiments. Figure 9.1 shows the number of composition and generalization operations performed per 50 episode intervals over the course of 6 iterations over the miniature language data from Experiments 3 and 4 with all learning operations enabled. As expected, there is a dramatic decline in the frequencies of these operations within the first iteration, and they slowly taper off in the remaining iterations as the learner settles on a grammar.

Figure 9.2 shows the same statistics for the Mandarin Chinese corpus from the "no decay" model in Experiment 1 and the "with decay" model in Experiment 2. Since neither experiment was able to run to the completion of the first iteration[43], the graphs show as many operations as the learner was able to perform. The number of composition operations shows a slight downward trend over time but the number of generalization operations is uncorrelated with the number of learning episodes, which is reasonable given that the Mandarin Chinese language is complex, the training data is sparse, the learner is conservative and is only on the first iteration of the data.

---

[43] due to an out-of-memory error during parsing, the "without decay" model only got halfway through the first iteration and the "with decay" model got through to about 80%.

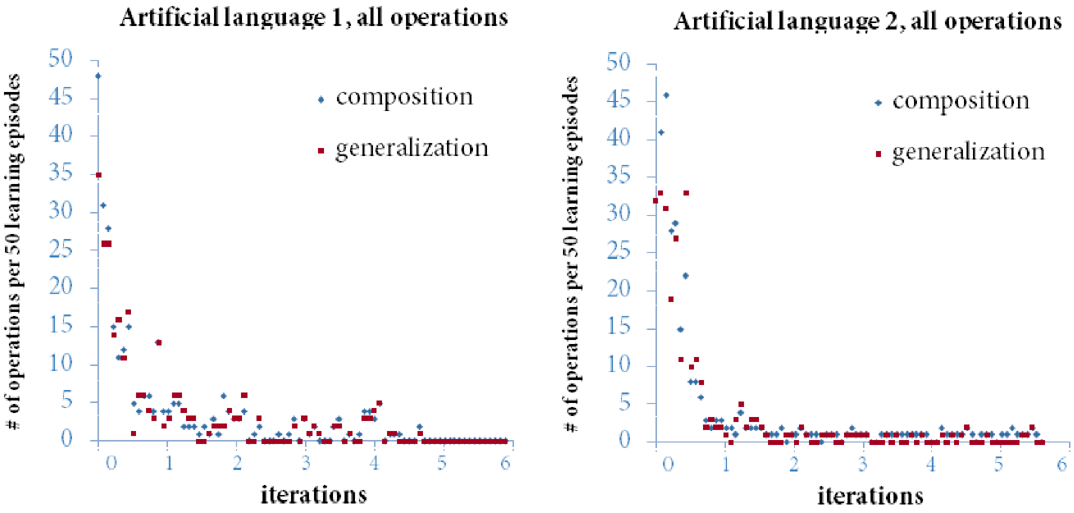

Figure 9.1 The number of composition and generalization operations performed by the learner decreases over time as the learner's grammar gets better and better at analyzing the learning input. Both graphs show the number of composition and generalization operations performed per 50 episode intervals over the course of 6 iterations over the training data with all learning operations enabled. The left shows data from one of the runs on the original artificial language; the right shows data from one of the runs on the modified (more complex) artificial language.

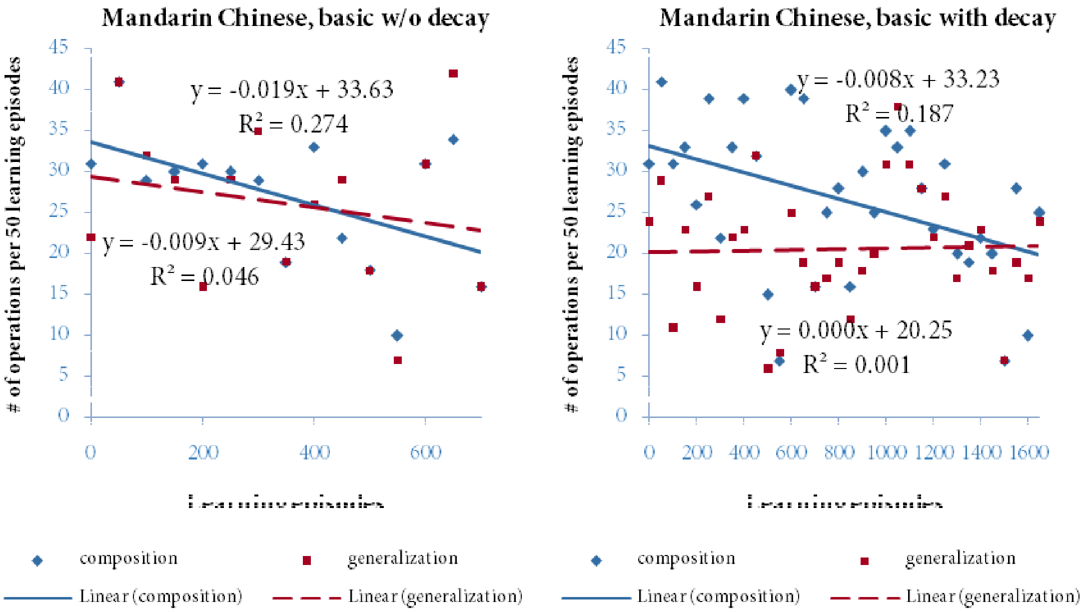

Figure 9.2 Both graphs show the number of composition and generalization operations taken per 50 episode intervals within the first iteration on the Mandarin Chinese corpus. The left shows data from the basic model without decay; the right shows data from the variation 1 model with decay. There is a slight downward trend for the number of composition operations but the number of generalization operations is uncorrelated with the number learning episodes (at least within the first iteration).

The quantitative results, on the other hand, are mixed: All variations of the model that had generalization capacity obtained satisfactory results on the unseen data in the miniature artificial languages in Experiments 3 and 4. In particular, Experiment 3 showed differentiation between combinations of learning operations on varying amounts of learning input. In Experiment 3, the refinement operations contributed more to the learner's ability to cope with unseen data when the amount of training data is limited, but the size of their effects seem to diminish when the miniature language gets more complicated in Experiment 4. On the other hand, in Experiments 1 and 2 the learned grammars from the naturalistic Mandarin Chinese data generalized somewhat to the unseen data but not particularly well.

The most likely reason why the model's success with miniature languages does not translate well to real languages is the scarcity of training data in comparison to the linguistic complexity of the corpus. By design, at 100% of the training data, one iteration of learning in Experiments 3 and 4 covered 80% of the 860 possible sentences allowed by the grammar. The ability of the learner to rapidly generalize based on experiencing many instances of the same construction is reflected in the steep drop-off in number of composition operations within the first iteration in Figure 9.1. (Do recall, however, that the performance of the learner degraded when the more complex miniature language in Experiment 4 was tested on the learner.)

By contrast, this kind of steep drop-off within the first iteration was not observed in Experiments 3 and 4, suggesting that the learner encounters data from different corners of the grammar throughout the first iteration. The refinement operations also inject into the grammar noise which normally goes away when the learner encounters more data. To be clear, some of the constructions proposed by these operations are correct, as evidenced by the learner's ability to achieve comparable quantitative results with far fewer constructions. However, sorting out the good refinements from the bad refinements is no easy task, and in the case of Experiments 1 and 2, there was not enough data to prune a sizable portion of the bad constructions and they ended up stealing probability mass away from the good constructions, causing the analyzer to return incorrect analyses.

9.2 Constructional generalization

There is no doubt that the ability to abstract away from the learning input is key to the linguistic productivity of a child learner. This helps her to both understand and produce utterances that she has never encountered before. As reviewed in Chapter 1, there is still a lot that is unknown about when and how children generalize. This dissertation is an attempt to lay out in precise terms some of the sources of information and computational processes that go into the formation of general constructions. One particular topic worth discussing is the ongoing competition between specific constructions and their generalizations, and the path the model takes to settle on the generalizations it makes.

Specific versus general constructions

How early children gain access to general argument structure constructions is a highly debated topic and consensus has yet to be reached in the language development community (Abbot-Smith et al., 2004; Akhtar & Tomasello, 1996; 1997; Conwell & Demuth, 2007; Fernandes, Marcus, Di Nubila & Vouloumanos, 2006; Fisher, 2002; Hirsh-Pasek, Golinkoff & Naigles, 1996b; Tomasello, 2000). There is agreement, however, that qualitative differences exist between what children are able to do with novel verbs in comprehension and production tasks at a younger and an older age. The current work contributes to this ongoing discussion in emphasizing that generalization is not an all-or-nothing process. The model presented here is constructed in such a way that lexically-specific constructions emerge before generalizations, but the pace and scope of generalization in the model is dependent on a number of factors. One such factor is demonstrated in variation II of the basic model (Section 7.2) where the statistic update discount factor @was set to 0.2 instead of 1.0. The pace of generalization slowed down dramatically since the learner, in a sense, did not trust the generalizations as much and was not eager to use them, which in turned caused many of the generalizations to be lost to decay.

Additionally, the scope of generalization vis-à-vis the size of constructional categories also grows larger over the course of learning. Of interest here is the notion of the constructional context of a construction =, loosely defined as the syntactic and semantic configuration of each construction that takes =as a constituent. For example category =may be a category of human nouns that has been created as a pre-verbal constituent which is connected semantically to the verb’s agent role. The category =is a “proto-subject”, so to speak, and we will call this configuration of form and meaning relations its constructional context. Another category >may be another category of human nouns and may even share a number of members with =, but if >is used as a post-verbal constituent connected to the main verb’s patient role (i.e. a “proto-object”), its constructional context is different from that of =’s. The constructional context of each category has the property of being mostly preserved through generalization. This is because constructional categories (i.e. abstract constructions) are by definition of ECG never instantiated on their own and always used by some other concrete constructions. The merging of two categories must be the result of their users being generalized, and the constructional context must therefore automatically be preserved.

There are only two scenarios in which a category =extends beyond its constructional context. The first is when more than one of its members is used in other constructions of a different constructional context and a generalization occurs over those constructions. That is, if, say, category =from above has as members WO3-N (I), NI3-N (you) and YI2-N (aunt), and there happens to be a generalization between GEI3-WO3 (give me) and GEI3-NI3 (give you), category = will be used as a constituent in the new GEI3- =construction. The resulting category =will have properties of both a proto-subject and a proto-object.

The other operation that violates constructional context distinctions in the current implementation is the category expansion operation. It explicitly looks only for semantic similarity between existing categories and other non-members, including other constructional categories.

The resulting generalization behavior is not incompatible with the idea of Radical Construction Grammar (Croft, 2001), where the notions of grammatical subjects and objects are not defined except with respect to the particular argument structure constructions that they are a part of. This kind of very conservative generalization is also compatible in spirit with children's learning behavior as demonstrated by Gerken's artificial language learning experiment where 9-month-old infants were given stimuli that can be explained by both a conservative and an aggressive generalization (Gerken, 2006). Specifically, when the exposure stimuli all ended in a particular syllable *di* while also obeying a general AAB pattern, infants chose the more conservative, syllable-based generalization.

Obviously, the number of possible generalizations in a real language is far greater than that in an artificial language and psychologists are only beginning to understand what facilitates a child's extension of existing verb-argument patterns to new verbs. There is evidence both of verbcentric generalizations (e.g. *want* object ,) (Tomasello, 2000) and nominal /morphology-anchored generalizations (e.g. *I'm* verb *ing it*) (Childers & Tomasello, 2001), putting the status of verb meaning as the driving force behind generalization in question (Ninio, 2005). The generalization mechanism in the model separates the issues of construction retrieval (i.e. which constructions are similar enough to the ones actively in use to perform generalization) from the actual act of generalization (i.e. how general should the new construction be — to put it in the context of Gerken's experiment — should it be the "end in *-di*" hypothesis or the AAB hypothesis). As explained in Chapter 4, both item-based and semantics-based strategies have been tested as the construction retrieval mechanism and the item-based strategy seemed to lead to generalizations that are too broad too quickly. However, this is exactly the kind of question that this learning model is designed to ask and make predictions for. Figure 9.3 gives a taste of the factors in the model that can influence how general the grammar becomes, and how quickly.

statistics	€ discount factor γ for the statistics updates for new constructions which reflects the level of confidence the learner has in its correctness
generalization	€ retrieval of constructions from grammar € additional criteria for initiating a generalization (e.g. relative frequencies of the two specific constructions) € whether specific constructions are kept after generalizations € how many "versions" of constructions to keep around if further generalizations are made
category merge	€ criteria for category merges (e.g. number or percentage of overlapping members, semantic distance) € whether users of the merged categories automatically have their contextual constraints relaxed at the same time
category expansion	€ criteria for extending a category (e.g. how many of the same category does the learner have to see to be confident about the extent of the category)

Figure 9.3 Examples of factors in the model that affect how quickly the learned grammar becomes general.

We have discussed how the discount factor γ and the construction retrieval strategy affect the shape of the grammar learning trajectory. Within the generalization operation, there are a few other operational details that affect learning. The first has to do with additional dimensions of comparison between a set of constructions for them to qualify as candidates for generalization. One such dimension that almost certainly matters is the relative frequencies of the constructions being generalized (Goldberg, Casenhiser & Sethuraman, 2004; Gomez, 2002; Hudson Kam & Newport, 2005; Thompson & Newport, 2007). Another operational detail that has big implications for the learning is closely related to the incremental nature of generalization in this model, and that has to do with what happens to the specific construction after it has been generalized over.

Bybee has argued based on phonological contractions that some high-frequency, lexically specific forms are retained in the grammar despite the availability of more general constructions (Bybee & Scheibman., 1999). Currently, a construction and its generalizations (and their subsequent generalizations) are all kept in the grammar. The basic idea is that these constructions will compete in usage and the ones at an inappropriate level of generalization eventually be purged due to decay. This implementation is certainly too naïve, especially in consideration that the amount of training data is severely limited. Thus the analyzer was likely to have been hampered by the amount of ambiguity introduced into the learned grammars in this process.

Bayesian learning approaches

Amongst the first places to look for solutions to properly model the competition between specific and general constructions is Bayesian learning approaches, which are certainly compatible with the current learning framework. The goal here is to find the most probable grammar $\hat{G}$ given a set of utterances U situated in contexts Z, which can be expressed as the product of the data likelihood and the grammar prior using Bayes rule and dropping the normalizing denominator,

$$\begin{aligned}\hat{G} &= \underset{G}{\operatorname{argmax}} P(G \mid U, Z) \\ &= \underset{G}{\operatorname{argmax}} P(U \mid G, Z) P(G \mid Z)\end{aligned}$$

and since the grammar does not depend on context, the grammar prior term can be simplified:

$$\hat{G} = \underset{G}{\operatorname{argmax}} P(U \mid G, Z) P(G)$$

By making an independence assumption between the utterances, the data likelihood can be estimated as the product of the probability of each utterance given the grammar and its context. By introducing a variable for the analysis of utterance and realizing that the utterance is deterministic given an analysis, the total data likelihood can be estimated as the product of the probabilities of all analyses.

$$\begin{aligned}P(U \mid G, Z) &= \underset{u}{} P(u \mid G, z) \\ &= \underset{a}{} P(u \mid a, G, z) P(a \mid G, z) \\ &= \underset{a}{} P(a \mid G, z)\end{aligned}$$

The probability of an analysis given a grammar G and a context z is exactly the probability that Bryant's best-fit analyzer estimates in its factored model (Bryant, 2008b). Given that, the data likelihood term is straightforward to calculate but difficult to implement in a cognitively plausible way — an accurate data likelihood term requires re-analysis of all previously encountered situated utterance, which, by its memory requirements alone is implausible for a child learner.

The even trickier bit, however, is to define a proper prior probability distribution for the grammar, $P(G)$. This is not at all straightforward for the desired qualities we want for a grammar, which is supposed to have some amount of redundancy between specific and general constructions. A grammar prior based on a simplicity measure has been attempted by Perfors, Tenenbaum, and Regier (personal communications) in their Bayesian selection of induced grammars , but the prior again does not capture the desire for specific constructions to co-exist with generalizations. A related information-theoretic approach, Minimum Description Length (MDL), has been used in Chang's model of construction grammar learning (Chang, 2008). MDL minimizes the total description length of the data and the grammar and is therefore designed to achieve the optimal level of compactness of the grammar. However, it suffers from the same criteria misfit as the Bayesian approach: if learning of new constructions are incremental such that all pieces of encountered data have been covered by some specific construction, if those specific constructions are not discarded after a generalization operation, and if encountered data (so far) is all that the learner has to go by in evaluating the description length, then there is always a net increase in description length after any generalization operation. Chang's model attempts to alleviate the problem by assuming that specific and general constructions share representational substrates and reducing the length of specific constructions that have been generalized.

At the end of the day, the ad-hoc nature of choosing a Bayesian grammar prior or a grammar length heuristic reflects a lack of understanding of the representations of and the interactions between abstract and specific grammatical knowledge in the human brain in the broader cognitive science and psycholinguistic communities. Until these grander challenges are met, more localized, limited applications of Bayesian learning principles can be explored in the learning model. There are some obvious parallels between the learning of grammatical categories (e.g. is the constituent after the word `CHI1` (eat) fillable by any word that refer to medicines, or food item, or physical object?) and the learning of linguistically-defined object categories (e.g. is *dog* a label for Dalmatians only, or for furry 4-legged animals, or for living things that run around?) (Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007) and it may be a fruitful direction to explore both in children and computational models.

9.3 Other kinds of constructions

Looking past the initial stage of combining content words (e.g. nouns, verbs, and the occasional directional particles that have image-schematic meanings), two particular kinds of constructions proved to be troublesome for the current learning model: constructions that have non-compositional meanings and function words. This section describes what the phenomena are, why they are difficult for the learner, and offers a sketch of new learning operations that helps cope with problems.

Constructions with non-compositional meaning

Non-compositional meaning refers to meaning components introduced into the new construction that cannot be attributed to any of its constituents. A typical example in English is the *What's X doing Y*? construction (Kay & Fillmore, 1999) whereby surprise and/or disapproval is expressed along with the question, as in *What's a nice girl like you doing in a place like this?*

Of course the learner model is not expected to learn a construction with pragmatics as complex as the WXDY construction right off the bat. However, there is a wide range of non-compositional meanings encoded by constructions. Some are in the physical motion domain, such as the caused motion construction in English, e.g. *he sneezed the napkin off the table,* or the serial verb construction in Chinese that encode sequences of motions, e.g. *guo4 lai2 chi1* (cross $\text{DIR}_{\text{towards}}$ eat / come over to eat). Some are in the temporal domain, such as a slightly different serial verb construction in Chinese that describe concurrent event, e.g. *zuo4 zhe chi1* (sit DUR eat

/ sit while you eat). A good number more are in the causal domain, such as the resultative constructions in English and Chinese such as *he drank himself silly* and *cha1 gan1+jing4* (wipe [it] clean) or ditransitive constructions such as *he baked her a cake* and *gei3 a1+yi2 chi1* (give aunt eat / give it to aunt for her to eat)). There are obviously also those like WXDY whose meaning is less concrete, such as the implied comparison in the *let alone* construction (Fillmore, Kay & O'Connor, 1988) as in *he can barely walk, let alone run a marathon*.

Learning any of these constructions requires the ability to construe the current scene as more than the sum of its parts by attributing physical, temporal, causal, or other relations to its components. The current learning framework does not facilitate the learner in any way by pre-segmenting the scenes with these relations. Instead, as described briefly in Section 4.1, the learner has to postulate a coherent meaning when multiple meaning roots are present in the new composition.

The mechanism in the current learner for selecting these relations is very crude and introduces quite some amount of ambiguity into the grammar. Compounded with the noise already present in the context-fitting process, the non-compositional meaning option in the composition operation hurt the learner's ability to analyze sentences correctly in the pilot runs.

This is in a way unsurprising. As the examples above illustrate, there is not a whole lot of syntactic distinction between the constructions that express motion sequence, concurrent motion, and resultative meanings. All of them basically manifest themselves as serial verb constructions with possibly an intervening aspect marker. Simulation is often required to properly differentiate the relations between the two events expressed by the serial verbs. For example, in *cha1 gan1+jing4* (wipe clean / wipe [it] clean), knowledge about wiping potentially causing a change of state will help to determine that the wipee may become clean as a result of the wiping process, otherwise the learner may pattern this after *guo4 lai2 chi1* (cross $\text{DIR}_{\text{towards}}$ eat / come over to eat), where the crosser is also the eater and the two processes are ordered temporally. This is exactly the kind of embodied knowledge that a learner ought to have access to; it is just that the current implementation of the simulation mechanism is not detailed enough to support such inference. A more fully developed learning model will make use of the context model and simulation mechanism to determine if a proposed non-compositional meaning is appropriate in context.

Function morphemes

Function morphemes, as the name suggests, are defined with respect to the relational functions they play in a construction. They are a bit more difficult to acquire because they are often unstressed, but the regularity in appearance of the obligatory function morphemes (such as the articles in English) are noted by children as young as 2 (Gerken & McIntosh, 1993). The closed-class nature of these function morphemes, as well as other cues that are regularly present in the input such as prosody, has also been argued to assist a child in the formation of phrasal groupings (Morgan, Meier & Newport, 1987; Morgan & Newport, 1981).

Function morphemes are learned in the current model only in an indirect way — the revision operation attempts to use collocating function morphemes (as well as content morphemes) to differentiate two conflicting constructions. The revision operation, unsurprisingly, turns out to be noise-prone. This subsection gives a sketch on how bigram probabilities can be exploited in this model to form "proto-construction" units that may reduce the need for subsequent revisions.

construction	gloss	sensible?
GE-HUAI4-c029	CLS - broken	
ZAI4-MO3-c047	again - apply	yes
RENG1-AO-SP-c065	throw - SFP	?
YONG4-CV-PING2ZI-c131	CV$_{\text{instrument}}$ - bottle	yes
GAN4MA2-WH-YA-SP-c219	how come - SFP	yes
KUAI4-CHI1-c377	quick - eat	yes
DA4GE4-DE-NOM-c379	big - NOM	yes
YI2-GE-c382	one - CLS	yes
MA1-GEI3-CV-c395	mother - CV$_{\text{benefactive}}$	
XIA4-DI4-c396	LOC$_{\text{downwards}}$ - ground	
HUI4-CHI1-c616	able - eat	yes
LIANG3-GE-c625	two - CLS	yes
QI3-LAI2-c757	rise - DIR$_{\text{up}}$	yes
WANR2-NE-SP-c787	play - SFP	?
NEI3-GE-c824	that - CLS	yes
GUAI3-GUO4-c849	turn - DIR$_{\text{across}}$	yes
DIAO4-LE-c984	drop - PFV	yes
NING3-ZHER4-c1050	twist - there	yes
ZHAO4-ZHE-c1296	mimic - DUR	yes
TIAO4-BA-SP-c1368	dance - SFP	yes
FU2-ZHE-c1404	support - DUR	yes

Figure 9.4 Examples of proto-constructions learned in a pilot run by chunking any bigrams between content morphemes and function morphemes that exceed 0.35 into a new construction.

The most simple-minded algorithm looks for bigrams between a content morpheme c and a function morpheme f that exceed a threshold, i.e. $P(c \mid f) >$ threshold or if $P(f \mid c) >$ threshold). Given these correlated units, the learner can create a new construction that has the content morpheme and the function morpheme as constituents and use the meaning pole of the content morpheme as the meaning of the new construction. In a pilot run where the bigram probability threshold was set to 0.35, the learner began to chunk content morphemes with collocated function morphemes, leading to the list of constructions learned in Figure 9.4. The figure shows on the leftmost column the proto-constructions, their glosses in the middle, and indicates whether each proto-construction forms a good unit (i.e. whether it is reasonable for the protoconstruction to be a constituent of some other construction). While two protoconstructions are questionable (in both cases the verb may be grouped prematurely with the sentence final particle), 16 of the 21 protoconstructions found are reasonable combinations.

This pilot run is a small proof of concept that bigram statistics can help discover phrasal units that may be helpful in anchoring the analysis of an utterance in the naturalistic Mandarin Chinese data. The idea of using statistics in the input is certainly not new. Thompson and Newport (2007) have conducted experiments where adults successfully learned artificial languages where the difference in word class transitional probabilities provide the only cues to the phrasal structure. Mintz (2003; 2006) has also demonstrated with CHILDES corpus data that distributional cues in the form of frequent frames are powerful tools for creating word classes. These and more sophisticated kinds of statistics are also the bread and butter of statistical NLP, and one of the key insights in Klein's constituent-context model (CCM) (2004) is the use of distributional cues along with a non-crossing bracketing constraint. As reviewed in Chapter 3, however, there is a disconnect between these kinds of statistically-derived phrase structures in induced grammars and the semantically-rich grammatical structures found in natural languages. This dissertation has focused on the formation of the latter using semantics as the primary source of information as well as the target for learning; it remains to be worked out how best to integrate statistically-derived structures such as these protoconstructions in the learning model.

9.4 Looking at language learning as a whole

Taking a step back, this dissertation addresses but a very small piece of the puzzle called language development. Many open questions remain; this section tackles some of the more pressing ones related to word learning, concept learning, morphosyntactic development, and real situational contexts.

Word learning

As alluded to many times throughout this dissertation, word learning is a process that is very much tied up with grammar learning developmentally and this model has made the arbitrary choice of starting with a set of known words and no knowledge of syntax. Undoubtedly, words are not learned in isolation from the rest of language. Verbs, in particular, offer particular construals of events and experience but the meaning of verbs, by their relational nature, are necessarily conflated with the rest of the scene and the process of teasing out the verb meanings involves generalizing cross-situationally over scene types as well as other the arguments they appear with. Verbs are therefore difficult to learn, as Gleitman and colleagues have shown in the human simulation experiments (Gillette et al., 1999; Gleitman, Cassidy, Nappa, Papafragou & Trueswell, 2005). The initial verbs in the models' grammar can be thought of as codified associations between linguistic forms, motor programs, and scenes, and can certainly be wrong in the beginning. The pace and scope of generalization is expected to be affected by the schema hierarchy in the following sense: for verbs that make fine-grain distinctions such as causality (e.g. knock over versus fall) or agentivity e.g. (trip versus fall), the danger of attributing too much knowledge to the initial learner lies in precluding generalizations that may otherwise be possible given fuzzier semantic definitions.

Undoubtedly, new words are constantly being learned throughout language development as well. This is further supported by evidence that syntactic development aids vocabulary development by allowing learners to infer the meaning of new verbs through the syntactic frame in which they occur (Fisher, 2002; Gleitman, 1990; Naigles, 1996). Also referred to as syntactic bootstrapping (Landau & Gleitman, 1985), this process can be a powerful mechanism in later language development and is recently found to play a role in the acquisition of a "worst-case" scenario language like Mandarin where argument omission is the norm (Lee & Naigles, 2008).

Ongoing word learning is theoretically compatible with the current framework. In addition to manual experiments with a gradually expanding vocabulary, new verbs are in theory learnable using the current model with a slight modification to the mechanism that learns non-compositional meaning. Specifically, if a novel action involving two entities is demonstrated in a sentence using a novel verb, say *blick*, the learner is left with multiple meaning components in the analysis (i.e. the mentioned entities) that it needs to relate with each other. Instead of trying to find some contextually-appropriate temporal or causal relations to explain the relations between events like in the non-compositional meaning case, here the learner can look to the situational context for events that involve the mentioned entities. Recognizing that the novel action not only relates the mentioned entities but also has an associated (novel) motor program, the learner can posit the motor program as the meaning poles of new compositions, leading to concrete constructions such as `YOU-BLICK-IT`, `I-BLICK-THIS`. Overtime, the learner will have a number of these contextually bound, lexically specific constructions, at which point the learner may generalize over them. The resulting general construction will have as constituents the novel verb and placeholders for its verb arguments and the associated motor program as its meaning. In the example, the resulting general construction, `CAT001-BLICK-CAT002`, contains all the lexical semantics of the novel verb *blick* plus semantic restrictions on its arguments, such as `Human` for the word preceding *blick* and `Physical_Object` for the word following *blick*. At this point the learner will have essentially learned the meaning of the new verb, though it will take another new operation to sub-analyze the `CAT001-BLICK-CAT002` construction in order to attribute lexical meaning to the verb directly.

Concept learning

Concepts are another domain that rapidly changes throughout development. Conceptual development is an issue that at first glance seems orthogonal to grammar development, but is upon closer examination intricately linked. There is a wealth of foundational work in the area of linguistic relativity that examines how language structures concepts, in particular in the domains of spatial concepts (Bowerman, 1996; Choi & Bowerman, 1991; Landau & Gleitman, 1985; Munnich, Landau & Dosher, 2001; Tversky & Lee, 1998) and color (Kay, Berlin, Maffi & Merrifield, 1997; Kay & Regier, 2006). This leads to another field of research on the Whorfian hypothesis (Whorf, 1956) which looks at how language influence thought. Specifically, linguistically structured concepts are found to influence thoughts even in non-linguistic tasks in various domains (Boroditsky, 2001; Drivonikou, Kay, Regier, Ivry, Gilbert, Franklin & Davies, 2007; Gilbert, Regier, Kay & Ivry, 2006; Winawer, Witthoft, Frank, Wu, Wade & Boroditsky, 2007).

Whorfian effects notwithstanding, a language learner must still learn the conceptual category distinctions dictated by the language. Luc Steels and colleagues have a series of models based on Fluid Construction Grammar (FCG) that model how language and concepts develop from a communication system and language evolution point of view (Steels, 2003; Steels, 2006; Steels & Version, 2004), but there is little computational work that focuses directly on how language and concepts co-develop in ontogeny. This is a challenging area of research, not least because the co-learning of the two domains is non-monotonic: changes in the conceptual system may inform the grammar that render existing constructions incorrect, and all the grammatical knowledge derived from those incorrect constructions now need to be revised.

Morphosyntactic development

This dissertation has focused primarily on the use word order and free function morphemes as indicators of semantic relations, ignoring inflectional morphology as a syntactic element. This was done partly out of convenience since Mandarin Chinese does not use inflectional morphology but also largely out of necessity since the available constructional analyzer system has no provision for morphology. However, current work is being done in the research group to interface the constructional analyzer with a morphological analyzer (see Section 9.1.6 of (Bryant, 2008a)). This has the added benefit of turning the current lexicalized analyzer into an unlexicalized one, which will greatly reduce the memory requirements of the analyzer and may even lead to some amounts of speed up.

Once the morphological capability of the analyzer is in place, the learner can be extended to use morphology as a form cue in the following way. The morphological analyzer decomposes the morphology into a constructional schema[44] containing features representing the morphological structure of each word, which in the beginning of learning may be as rudimentary as the form of the morpheme. With some amount of hand waving, we can imagine that these morphological features are stored as constructional features in concrete constructions created through the composition operation and are generalized just like meaning schemas through the

---

[44] It has not been mentioned in earlier chapters since it was not necessary, but the constructional pole as well as the form pole of a construction can be typed just in the same way as the meaning pole. Form schemas and constructional schemas can be defined and they are treated with the same exact unification semantics as meaning schemas.

generalization operation. At this point, the same sort of sub-analyzing operation as mentioned in the word learning section will be able to split the general construction and attribute functions to individual morphemes.

Real situational contexts

This section is titled “real situational contexts” because the situational context that the current model relies on is not only symbolically represented but also drastically simplified. The problem of scaling a language learning model to the real kind of messy situational contexts that a child learns in is essentially AI-complete and requires computational sophistications with vision systems, speech recognition systems. Roy and colleagues are tackling some of these challenges with vision-enabled robots with some success on word learning and very elementary syntax (Gorniak & Roy, 2007; Roy, 2002; Roy, 2003).

Vision and speech recognition systems not withstanding, there are still some grand challenges in modeling contexts in any real sense. Here are some observations about the difficulty of the task from working on this particular learning model and child language data:

€ Metonymy and construal is everywhere in child language interaction. Consider scenarios where parent and child are engaged in story time. Picture books showing pictures of cars are present in the same scene as toy cars and real cars. The same words can refer to the pictures of the cars, the toy cars in the room, the real cars sitting in the driveway, or even in some cases, the physical sheet of paper on which the pictures are printed. It is no easy task for a computational system to see a word *car* and try to resolve its intended referent.

€ Scenes are always perspectivized and so is the language describing the scenes. Verbs like *give* and *receive* impose perspectives on the scene, but so do locative words with a reference object such as *inside*, *outside*, *here*, or *there*. Properly representing the meaning of these locative words requires even richer semantics and a context model that is capable of representing the physical properties of entities.

€ As it turned out, figuring out whether an utterance describes a past, future or irrealis event without having any knowledge of tense aspect marking was very difficult for the learning model. The annotated speech-acts, which were inferred from intonation, were not an entirely reliable indicator of when (if at all) in the situational context a mentioned event takes place: A *requesting-action* utterance can be a pre-emptive request for a child to not do something, or for the child to stop doing something she's doing. An *explaining* utterance can be a declaration of an intention to do something or a description of what the speaker has just done. Even in an *admonishment*, parents often threaten the child with some future action if the child continues to do something she has been doing. This difficulty with resolving events to context led to a sizable amount of noise in the current model and is a difficulty that a truly situated model of language learning and use must overcome.

9.5 Summary

The model of early grammar learning presented in this dissertation benefits from bootstrapping from situational context as well as the richness of semantic knowledge available to the learner. It represents a first step in setting up a precise computational experiment framework with explicit operational definitions of learning processes and clearly defined sources of knowledge. Model parameters are easily adjustable for computational experiments, as demonstrated, and we believe that a combination of learning experiments with real and artificial language will prove fruitful for understanding the process of language learning.

Bibliography

Abbot-Smith, K., Lieven, E. & Tomasello, M. (2004). Training 2;6-year-olds to produce the transitive construction: the role of frequency, semantic similarity and shared syntactic distribution. Developmental Science 7, 48-55.

Akhtar, N. & Tomasello, M. (1996). Two-year-olds learn words for absent objects and actions. British Journal of Developmental Psychology 14, 79-93.

Akhtar, N. & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. Developmental Psychology 33, 952–65.

Alishahi, A. (2008). A probabilistic model of early argument structure acquisition. Ph.D. Dissertation, University of Toronto.

Alishahi, A. & Stevenson, S. (2008). A Probabilistic Model of Early Argument Structure Acquisition. Cognitive Science 32, 789-834.

Altmann, G. T. M. & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. Journal of Memory and Language 57, 502-18.

Altmann, G. T. M. & Kamide, Y. ((under review)). Discourse-mediation of the mapping between language and the visual world: eye-movements and mental representation.

Aslin, R. N., Saffran, J. R. & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. Psychological Science 9, 321-4.

Bailey, D. R., Feldman, J. A., Narayanan, S. & Lakoff, G. (1997). Modeling Embodied Lexical Development. 19th Cognitive Science Society Conference.

Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998). The Berkeley FrameNet project. 1998 International Conference on Computational Linguistics.

Bates, E., Chen, S., Tzeng, O., Li, P. & Opie, M. (1991). The Noun-Verb problem in Chinese aphasia. Brain and Language 41, 203-33.

Bergen, B. K. & Chang, N. (2005). Embodied Construction Grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (eds.), Construction Grammars: Cognitive Groundings and Theoretical Extensions. Philadelphia, PA: John Benjamins.

Bergen, B. K., Chang, N. & Narayan, S. (2004). Simulated Action in an Embodied Construction Grammar. Proc. 26th Cognitive Science Society Conference.

Bergen, B. K., Lindsay, S., Matlock, T. & Narayanan, S. (2007). Spatial and Linguistic Aspects of Visual Imagery in Sentence Comprehension. Cognitive Science 31.

Bergen, B. K. & Wheeler, K. B. (2005). Sentence Understanding Engages Motor Processes. Twenty-Seventh Annual Conference of the Cognitive Science Society.

Bloom, P. (2002). How Children Learn the Meaning of Words. Cambridge: MIT Press.

Boroditsky, L. (2001). Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time. Cognitive Psychology 43, 1-22.

Bowerman, M. (1996). Learning how to structure space for language: A crosslinguistic perspective. Language and space, 385-436.

Brandone, A. C., Pence, K. L., Golinkoff, R. M. & Hirsh-Pasek, K. (2007). Action Speaks Louder Than Words: Young Children Differentially Weight Perceptual, Social, and Linguistic Cues to Learn Verbs. Child Development 78, 1322-42.

Bryant, J. (2008a). Best-Fit Constructional Analysis. Ph.D. Dissertation, University of California, Berkeley.

Bryant, J. (2008b). Exploiting statistical information in constructional analysis. Paper presented at the Fifth International Conference on Construction Grammar.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V. & Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. Cognitive Brain Research 24, 355-63.

Burke, M., Lam, O., Cahill, A., Chan, R., O'Donovan, R., Bodomo, A., Genabith, J. v. & Way, A. (2004). Treebank-based acquisition of a Chinese Lexical-Functional Grammar. 18th Pacific Asia Conference on Language, Information and Computation.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. Language 82, 711.

Bybee, J. & Scheibman., J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. Linguistics 37, 575-96.

Cahill, A., Burke, M., O'Donovan, R., Riezler, S., van Genabith, J. & Way, A. (2008). Wide-Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation. Computational Linguistics 34, 81-124.

Casenhiser, D. & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. Developmental Science 8, 500-8.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H. & Carlson, G. N. (2002). Circumscribing Referential Domains during Real-Time Language Comprehension. Journal of Memory and Language 47, 30-49.

Chambers, K. E., Onishi, K. H. & Fisher, C. L. (2003). Infants learn phonotactic regularities from brief auditory experience. Cognition 87, B69-B77.

Chang, N. (2008). Constructing grammar: A computational model of the emergence of early constructions. Ph.D. Dissertation, University of California Berkeley.

Chang, N., Feldman, J. & Narayanan, S. (2004). Structured Connectionist Models of Language, Cognition and Action. Ninth Neural Computation and Psychology Workshop. Plymouth, UK.

Chang, N. & Gurevich, O. (2004). Context-driven construction learning. 26th Annual Meeting of the Cognitive Science Society. Chicago, IL.

Chang, N. & Mok, E. (2006a). Putting Context in Constructions. Paper presented at the The Fourth International Conference on Construction Grammar (ICCG4).

Chang, N. & Mok, E. (2006b). A Structured Context Model for Grammar Learning. The 2006 International Joint Conference on Neural Networks. Vancouver, BC.

Chen, J., Bangalore, S. & Vijay-Shanker, K. (2005). Automated extraction of Tree-Adjoining Grammars from treebanks. Natural Language Engineering 12, 251-99.

Childers, J. B. & Tomasello, M. (2001). The Role of Pronouns in Young Children's Acquisition of the English Transitive Construction. Developmental Psychology 37, 739-48.

Choi, S. & Bowerman, M. (1991). Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. Cognition 41, 83-121.

Clark, A. S. (2001). Unsupervised Language Acquisition: Theory and Practice. Ph.D. Dissertation, University of Sussex.

Conwell, E. & Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. Cognition 103, 163-79.

Croft, W. (2001). Radical Construction Grammar: syntactic theory in typological perspective. Oxford: Oxford University Press.

Dore, J. (1974). A pragmatic description of early language development. Journal of Psycholinguistic Research 3, 343-50.

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A. & Davies, I. R. L. (2007). Further evidence that Whorfian effects are stronger in the right visual field than the left. Proceedings of the National Academy of Sciences 104, 1097-102.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1997). Rethinking Innateness: A Connectionist Perspective on Development (Neural Networks and Connectionist Modeling). The MIT Press.

Erbaugh, M. S. (1992). The acquisition of Mandarin. In D. I. Slobin (ed.), The Cross-linguistic Study of Language Acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Feldman, J., Dodge, E. & Bryant, J. (to appear). A Neural Theory of Language and Embodied Construction Grammar. In B. Heine & H. Narrog (eds.), The Oxford Handbook of Linguistic Analysis.

Feldman, J. A. (2006). From Molecule to Metaphor: A Neural Theory of Language. Cambridge, MA: MIT Press.

Fernandes, K. J., Marcus, G. F., Di Nubila, J. A. & Vouloumanos, A. (2006). From semantics to syntax and back again: Argument structure in the third year of life. Cognition 100, B10-B20.

Fillmore, C. J., Kay, P. & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. Language 64, 501-38.

Fiser, J. z. & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences of the United States of America 99, 15822-6.

Fisher, C. L. (2002). Structural limits on verb mapping: the role of abstract structure in 2.5-year-olds' interpretations of novel verbs. Developmental Science 5, 55-64.

Gahl, S. & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. Language 80, 748-75.

Gahl, S. & Garnsey, S. M. (2006). Knowledge of grammar includes knowledge of syntactic probabilities. Language 82, 405-10.

Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain 119, 593-609.

Gallese, V. & Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. Cognitive Neuropsychology 22, 455-79.

Gentner, D. & Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist 52, 45-56.

Gentner, D. & Namy, L. L. (2006). Analogical Processes in Language Learning. Current Directions in Psychological Science 15, 297-301.

Gerken, L. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. Cognition 98, B67-B74.

Gerken, L. A. & McIntosh, B. J. (1993). Interplay of Function Morphemes and Prosody in Early Language. Developmental Psychology 29, 448-57.

Gilbert, A. L., Regier, T., Kay, P. & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. Proceedings of the National Academy of Sciences of the United States of America 103, 489-94.

Gillette, J., Gleitman, H., Gleitman, L. & Lederer, A. (1999). Human simulations of vocabulary learning. Cognition 73, 135-76.

Givón, T. (2001). Syntax: an introduction. Vol. 1. J. Benjamins.

Gleitman, L. (1990). The Structural Sources of Verb Meanings. Language Acquisition 1, 3-55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A. & Trueswell, J. C. (2005). Hard Words. Language Learning and Development 1, 23-64.

Gold, E. M. (1967). Language Identification in the Limit. Information and Control 10, 447-74.

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. Chicago: University of Chicago Press.

Goldberg, A. E., Casenhiser, D. & Sethuraman, N. (2004). Learning Argument Structure Generalizations. Cognitive Linguistics 15, 289-316.

Golinkoff, R. M., Hirsh-Pasek, K., Mervis, C. B., Frawley, W. B. & Parillo, M. (1995). Lexical principles can be extended to the acquisition of verbs. Beyond names for things: Young children's acquisition of verbs, 185-221.

Gomez, R. (2002). Variability and detection of invariant structure. Psychological Science 13, 431-6.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. Psychological Review 111, 3-32.

Gorniak, P. & Roy, D. (2007). Extended Article: Situated Language Understanding as Filtering Perceived Affordances. Cognitive Science 31, 197-231.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.), Syntax and Semantics, 3: Speech Acts. New York: Academic Press.

Heibeck, T. H. (1985). Word Learning in Children: An Examination of Fast Mapping. Paper presented at the Biennial Meeting of the Society for Research in Child Development.

Hirsh-Pasek, K., Golinkoff, R. & Naigles, L. (1996a). Single-Word Speakers' Comprehension of Word Order. The origins of grammar, 99-122.

Hirsh-Pasek, K., Golinkoff, R. & Naigles, L. (1996b). Young children's use of syntactic frames to derive meaning. The origins of grammar, 123-58.

Hockenmaier, J. & Steedman, M. (2002). Acquiring compact lexicalized grammars from a cleaner treebank. The International Conference on Language Resources and Evaluation (LREC). Las Palmas, Spain.

Hood, B. M., Willen, J. D. & Driver, J. (1998). Adult's Eyes Trigger Shifts of Visual Attention in Human Infants. Psychological Science 9, 131-4.

Horning, J. J. (1969). A study of grammatical inference. PhD Dissertation, Stanford University.

Hudson Kam, C. L. & Newport, E. L. (2005). Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. Language Learning and Development 1, 151-95.

Johnson, M. L. (1987). The body in the mind: The bodily basis of meaning, imagination, and reasoning. Chicago: University of Chicago Press.

Jusczyk, P. W. (1997). Finding and Remembering Words: Some Beginnings by English-Learning Infants. Current Directions in Psychological Science 6, 170-4.

Kay, P., Berlin, B., Maffi, L. & Merrifield, W. (1997). Color Naming Across Languages. Color Categories in Thought and Language.

Kay, P. & Fillmore, C. (1999). Grammatical constructions and linguistic generalizations: the What's X doing Y? construction. Language 75, 1-33.

Kay, P. & Regier, T. (2006). Language, thought and color: recent developments. TRENDS in Cognitive Sciences 10.

Kim, Y.-J. (2000). Subject/object drop in the acquisition of Korean: A cross-linguistic comparison. Journal of East Asian Linguistics 9, 325-51.

Kingsbury, P. & Palmer, M. (2002). From Treebank to Propbank. 3rd International Conference on Language Resources and Evaluation.

Kirkham, N. Z., Slemmer, J. A. & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition 83, B35-B42.

Klein, D. (2005). The Unsupervised Learning of Natural Language Structure. Ph.D. Thesis Dissertation, Stanford University.

Klein, D. & Manning, C. D. (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. Association for Computational Linguistics. Morristown, NJ, USA.

Lakoff, G. (1987). Women, Fire, and Dangerous Things. Chicago: University of Chicago Press.

Lakoff, G. & Johnson, M. (1980). Metaphors We Live By. Chicago: University of Chicago Press.

Landau, B. & Gleitman, L. R. (1985). Language and Experience: Evidence from the Blind Child. Harvard University Press.

Langacker, R. W. (1990). Concept, Image, Symbol: The Cognitive Basis of Grammar. Berlin, New York: Mouton de Gruyter.

Langley, P. & Stromsten, S. (2000). Learning Context-Free Grammars with a Simplicity Bias. Lecture Notes in Computer Science, 220-8.

Lee, H.-T. (1996). Theoretical issues in language development and Chinese child language. In J. C.-T. Huang & A. Li (eds.), New Horizons in Chinese Linguistics. Dordrecht: Kluwer.

Lee, J. N. & Naigles, L. R. (2008). Mandarin learners use syntactic bootstrapping in verb acquisition. Cognition 106, 1028-37.

Li, C. N. & Thompson, S. A. (1981). Mandarin Chinese: A Functional Reference Grammar. Berkeley, Los Angeles: University of California Press.

Li, P., Jin, Z. & Tan, L. H. (2004). Neural representations of nouns and verbs in Chinese: an fMRI study. Neuroimage 21, 1533-41.

Li, W. (2004). Topic chains in Chinese discourse. Discourse Processes 27, 25-45.

Ma, W., Golinkoff, R. M., Hirsh-Pasek, K., McDonough, C. & Tardif, T. (2008). Imageability predicts the age of acquisition of verbs in Chinese children. Journal of Child Language, 1-19.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Macwhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. Journal of Child Language 31, 883-914.

Maguire, M. J., Hennon, E. A., Hirsh-Pasek, K., Golinkoff, R. M., Slutzky, C. B. & Sootsman, J. (2001). Mapping words to actions and events: How do 18-month-olds learn a verb. Boston University Annual Conference on Language Development.

Mandler, J. (1992). How to Build a Baby: II. Conceptual Primitives. Psychological Review 99, 587-604.

Marcus, G. F., Vijayan, S., Bandi Rao, S. & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. Science 283, 77.

Markson, L. & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. Nature 385, 813-5.

Matlock, T. (2004). Fictive motion as cognitive simulation. Memory & Cognition 32, 1389-400.

Maye, J., Werker, J. F. & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82, 101-11.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. Cognition 90, 91-117.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. Action meets word: How children learn verbs, 31-63.

Morgan, J. L., Meier, R. P. & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of intonational and morphological marking of phrases to the acquisition of language. Cognitive Psychology 19, 498-550.

Morgan, J. L. & Newport, E. L. (1981). The Role of Constituent Structure in the Induction of an Artificial Language. Journal of Verbal Learning and Verbal Behaviour New York, N. Y. 20, 67-85.

Munakata, Y., McClelland, J. L., Johnson, M. H. & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. Psychological Review 104, 686-713.

Munnich, E., Landau, B. & Dosher, B. A. (2001). Spatial language and spatial representation: a cross-linguistic comparison. Cognition 81, 171-208.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V. & Rizzolatti, G. (1997). Object Representation in the Ventral Premotor Cortex (Area F5) of the Monkey. Journal of Neurophysiology 78, 2226-30.

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. Cognition 58, 221-51.

Narayanan, S. (1997). KARMA: Knowledge-based Action Representations for Metaphor and Aspect. Ph.D. dissertation Dissertation, University of California Berkeley.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. National Conference on Artificial Intelligence.

Newport, E. L. & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. Cognitive Psychology 48, 127-62.

Ninio, A. (2005). Testing the role of semantic similarity in syntactic development. Journal of Child Language 32, 35-61.

Perfors, A. (2008). Learnability, representation, and language: A Bayesian approach. PhD Dissertation, Massachusetts Institute of Technology.

Perfors, A., Tenenbaum, J. & Regier, T. (2006). Poverty of the stimulus? A rational approach. the 28th Annual Conference of the Cognitive Science Society.

Pullum, G. K. & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. The Linguistic Review 18, 9-50.

Quine, W. V. O. (1960). Word and Object. Cambridge, MA: Harvard University Press.

Roy, D. (2002). Learning Words and Syntax for a Visual Description Task. Computer Speech and Language 16.

Roy, D. (2003). Grounded spoken language acquisition: experiments in word learning. Multimedia, IEEE Transactions on 5, 197-209.

Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F. & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. Cognition 107, 479-500.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. Cognition 81, 149-69.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science 274, 1926-8.

Saffran, J. R., Pollak, S. D., Seibel, R. L. & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. Cognition 105, 669-80.

Saffran, J. R. & Thiessen, E. D. (2003). Pattern induction by infant language learners. Developmental Psychology 39, 484-94.

Shi, D. (2000). Topic and topic-comment constructions in Mandarin Chinese. Language 76, 383-408.

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C. & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. Brain and Language 101, 260-77.

Slobin, D. I. (1986). Crosslinguistic Evidence for the Language-making Capacity. In D. I. Slobin (ed.), The Crosslinguistic Study of Language Acquisition, vol. 2. Hillsdale, NJ: Lawrence Erlbaum Associates.

Snedeker, J., Li, P. & Yuan, S. (2003). Cross-Cultural Differences in the Input to Early Word Learning. Twenty-fifth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates.

Spivey, M. (2008). The Continuity of Mind. Oxford University Press.

Spivey, M. J. & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. Current Directions in Psychological Science 15, 207-11.

Steels, L. (2003). Evolving grounded communication for robots. Trends in Cognitive Science 7, 308-12.

Steels, L. (2006). Experiments on the emergence of human communication. TRENDS in Cognitive Sciences 10, 347-9.

Steels, L. & Version, D. (2004). Social and Cultural Learning in the Evolution of Human Communication. Evolution of Communication Systems: A Comparative Approach.

Talmy, L. (2000). Toward a Cognitive Semantics. Cambridge, MA: MIT Press.

Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation Dissertation, Yale University.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. Developmental Psychology 32, 492-504.

Tardif, T. (2006). But are they really verbs? Chinese words for action. Action meets word: How children learn verbs, 477-98.

Tardif, T., Shatz, M. & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. Journal of Child Language 24, 535-65.

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences 24, 629-40.

Thompson, S. P. & Newport, E. L. (2007). Statistical Learning of Syntax: The Role of Transitional Probability. Language Learning and Development 3, 1-42.

Tomasello, M. (1999). The Cultural Origins of Human Cognition. Cambridgem MA: Harvard University Press.

Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition 74, 209-53.

Tomasello, M. (2001). Perceiving intention and learning words in the second year of life. In M. Bowerman & S. Levison (eds.), Language Acquisition and Conceptual Development. Cambridge: Cambridge University Press.

Tomasello, M. (2003). Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press.

Tversky, B. & Lee, P. U. (1998). How space structures language. Spatial Cognition. An interdisciplinary approach to representing and processing spatial knowledge, 157-75.

Wang, Q., Lillo-Martin, D., Best, C. T. & Levitt, A. (1992). Null subject versus null object: Some evidence from the acquisition of Chinese and English. Language Acquisition 2, 221-54.

Whorf, B. L. (1956). Language, thought, and reality. MIT Press.

Wilcox, T. & Schweinle, A. (2002). Object individuation and event mapping: Developmental changes in infants' use of featural information. Developmental Science 5, 132-50.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R. & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy of Sciences 104, 7780-5.

Wolff, J. G. (1988). Learning syntax and meanings through optimization and distributional analysis. Categories and Processes in Language Acquisition, 179-215.

Wonnacott, E., Newport, E. L. & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. Cognitive Psychology 56, 165-209.

Xia, F., Han, C., Palmer, M. & Joshi, A. (2001). Automatically Extracting and Comparing Lexicalized Grammars for Different Languages. the Seventh International Joint Conference on Artificial Intelligence (IJCAI-2001). Seattle, Washington.

Xu, F. & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference. Psychological Review 114, 245.

Yang, F.-P. G. (2007). Conditional Constructions in Mandarin: A Neural Explanation. Ph.D Dissertation, University of California, Berkeley.

Zettlemoyer, L. S. & Collins, M. (2007). Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Appendix A.

A context-free representation of the ECG syntax

(reproduced from Bryant (2008))

ECGL → ECGL Schema | ECGL Cxn | ϵ

CxnKind → GENERAL CONSTRUCTION | CONSTRUCTION

Cxn → CxnKind IDENT ParentLOpt CBlockOpt FBlockOpt MBlockOpt

BlockType → : Typespec | ϵ

CBlockOpt → CONSTRUCTIONAL BlockType ConstitsLOpt ConstraintLOpt | ϵ

FBlockOpt → FORM BlockType ConstraintLOpt | ϵ

MBlockOpt → MEANING BlockType EvokedLOpt RolesLOpt ConstraintLOpt | ϵ

SchemaKinds → FEATURE SCHEMA | SEMANTIC SCHEMA | SCHEMA

Schema → SchemaKinds IDENT ParentLOpt EvokedLOpt …
RolesLOpt ConstraintLOpt

ParentLOpt → ParentL | ϵ

SubcaseOf → SUBCASE | SUBCASE OF

ParentL → ParentL , IDENT | SubcaseOf IDENT

Typespec → IDENT | EXTERNALTYPE

EvokedElement → EVOKES Typespec AS IDENT

EvokedLOpt → EvokedLOpt EvokedElement | ϵ

Role → IDENT OptType

OptType → : Typespec | ϵ

RolesLOpt → RolesL | ϵ

RolesL → RolesL Role | ROLES

Constit → OPTIONAL IDENT : IDENT OptConstitAnno
| EXTRAPOSED IDENT : IDENT OptConstitAnno
| IDENT : IDENT OptConstitAnno

OptConstitAnno → [ProbL] | ϵ

ProbL → PROB | PROB , PROB

ConstitsLOpt → ConstitsL | ϵ

ConstitsL → ConstitsL Constit | CONSTITUENTS

ConstraintLOpt → ConstraintL | ϵ

ConstraintL → ConstraintL OptIgnore Constraint | CONSTRAINTS

ChainOperator → ⟷ | BEFORE | MEETS

OptIgnore → IGNORE | ϵ

Var → SLOTCHAIN | IDENT

Constraint → Var ChainOperator Var | Var ⟵ IdentOrStr

IdentOrStr → EXTERNALTYPE | IDENT | STR

IDENT → [A-Za-z][0-9a-zA-Z-_]*

SLOTCHAIN → (IDENT.)+IDENT

EXTERNALTYPE → @[0-9a-zA-Z.-_]+

STR → “(\” | [∧\n”] | \{WHITE_SPACE_CHAR}+\)*”

PROB → .[0-9]+ | 1.0 | 1

Appendix B.

An annotated CHILDES transcript sample in XML

```
<?xml version="1.0" encoding="UTF-8"?>
<CHAT xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.talkbank.org/ns/talkbank"
xsi:schemaLocation="http://www.talkbank.org/ns/talkbank talkbank.xsd"
Version="1.1.3" Lang="zh" Corpus="beijing" Id="cx2" Date="1984-01-01">

  <Participants>
    <participant id="MOT" role="Mother" language="zh" />
    <participant id="CHI" role="Target_Child" language="zh" />
    <participant id="FAT" role="Father" language="zh" />
    <participant id="INV" role="Investigator" language="zh" />
    <participant id="UNC" name="Unclear" role="Unidentified" language="zh" />
  </Participants>

  <Setting>
    <entity cat="Livingroom" id="livingroom"/>
    <entity cat="Peach" id="peach"/>
  </Setting>

  <Setup>
    <binding field="location" source_ref="MOT" ref="livingroom"/>
    <binding field="location" source_ref="CHI" ref="livingroom"/>
    <binding field="location" source_ref="INV" ref="livingroom"/>
    <binding field="location" source_ref="peach" ref="coffeetable(livingroom)"/>
  </Setup>

  <event cat="Fetch" id="fetch01">
    <binding field="fetcher" ref="CHI"/>
    <binding field="fetched" ref="peach"/>
  </event>

  <u who="MOT" id="149">
    <clause>
      <w>ni3</w><w>rang4</w><wn><w>a1</w><wk type="cmp" /><w>yi2</w></wn>
      <w>chi1</w><t type="p" />
      <a type="speech act">
        <sa cat="requesting-action" id="u149sa1">
          <binding field="speaker" ref="MOT"/>
          <binding field="addressee" ref="CHI"/>
          <binding field="forcefulness" value="Normal"/>
        </sa>
      </a>
      <a type="vernacular">          你讓阿姨吃</a>
      <a type="gold standard">
        <semantic>
          <temporal_element left="1" right="2" cat="Permit" id="u149te1">
                  <binding field="permiter" left="0" right="1" ref="CHI"/>
                  <binding field="permitee" left="2" right="4" ref="INV"/>
                  <binding field="permitted" left="4" right="5" ref="u149ts2"/>
```

```
        </temporal_element>
        <temporal_structure left="0" right="5" cat="Ditransitive_Action"
        profiled="u149te1" id="u149ts1">
                <binding field="giver" left="0" right="1" ref="CHI"/>
                <binding field="recipient" left="2" right="4" ref="INV"/>
                <binding field="theme" left="4" right="5" ref="u149ts2"/>
        </temporal_structure>

        <temporal_element left="4" right="5" cat="Eat" id="u149te2">
                <binding field="eater" left="2" right="4" ref="INV"/>
                <binding field="food" ref="peach"/>
        </temporal_element>
        <temporal_structure left="2" right="5" cat="Transitive_Action"
        profiled="u149te2" id="u149ts2">
                <binding field="agent" left="2" right="4" ref="INV"/>
                <binding field="patient" ref="peach"/>
        </temporal_structure>
      </semantic>
    </a>
  </clause>
</u>

<event cat="Offer" id="offer02">
  <binding field="offerer" ref="CHI"/>
  <binding field="offeree" ref="INV"/>
  <binding field="offered" ref="peach"/>
</event>

<u who="MOT" id="150">
  <clause>
    <w>ni3</w><w>gei3</w><w>yi2</w><t type="p" />
    <a type="vernacular">        你給姨</a>
    <a type="speech act">
      <sa cat="requesting-action" id="u150sa1">
        <binding field="speaker" ref="MOT"/>
        <binding field="addressee" ref="CHI"/>
        <binding field="forcefulness" value="Normal"/>
      </sa>
    </a>
    <a type="gold standard">
      <semantic>
        <temporal_element left="1" right="2" cat="Give" id="u150te1">
                <binding field="giver" left="0" right="1" ref="CHI"/>
                <binding field="recipient" left="2" right="3" ref="INV"/>
                <binding field="theme" ref="peach"/>
        </temporal_element>
        <temporal_structure left="0" right="3"  cat="Ditransitive_Action"
        profiled="u150te1">
                <binding field="giver" left="0" right="1" ref="CHI"/>
                <binding field="recipient" left="2" right="3" ref="INV"/>
                <binding field="theme" ref="peach"/>
        </temporal_structure>
      </semantic>
    </a>
  </clause>
</u>

<event cat="Give" id="give03">
  <binding field="giver" ref="CHI"/>
  <binding field="recipient" ref="INV"/>
  <binding field="theme" ref="peach"/>
</event>
```

```
  <u who="INV" id="153">
    <clause>
      <wn><w>xie4</w><wk type="cmp" /><w>xie4</w></wn><t type="p" />
      <a type="speech act">
        <sa cat="answering" id="u153sa1">
          <binding field="speaker" ref="INV"/>
          <binding field="addressee" ref="CHI"/>
          <binding field="forcefulness" value="Normal"/>
        </sa>
      </a>
      <a type="vernacular">          謝謝</a>
      <a type="gold standard">
        <semantic>
          <temporal_structure cat="None"/>
        </semantic>
      </a>
    </clause>
  </u>


</CHAT>
```