Contextual Bootstrapping for Grammar Learning

Eva H. Mok

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-12
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-12.html

January 26, 2009

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.

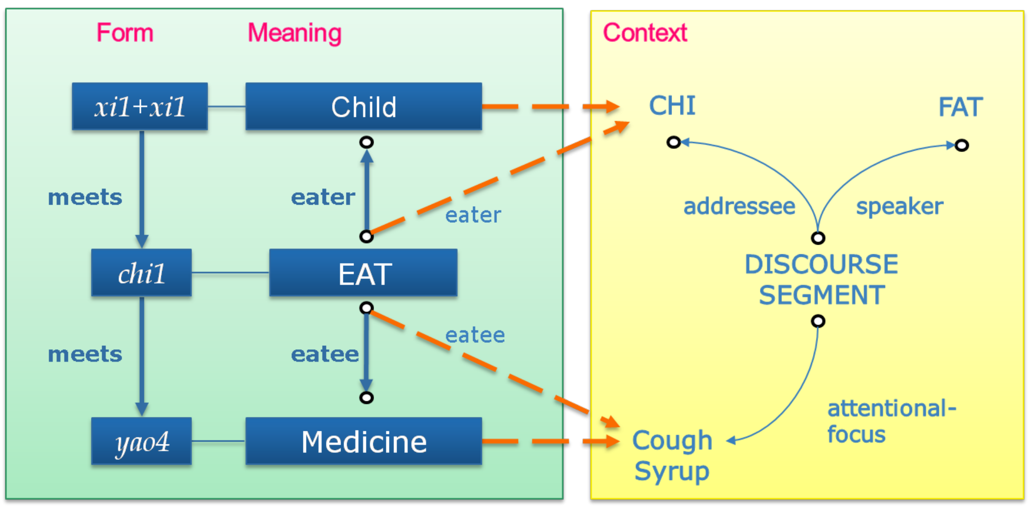

Figure 1.1 Interpreting the sentence *xi1xi1 chi1 yao4* (Xixi eat medicine) involves establishing constituency relations between the words and semantic relations between the corresponding meanings.

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 beginning are subcases of MORPHEME; it has no knowledge of any actual PHRASEs or CLAUSEs but will build them up over the course of learning. Extending from the idea that early constructions are structured around prototype scenes (Slobin, 1986), a clause is any construction that describes an event while a phrase can be anything else.

As illustrated in Figure 1.2, the system implements a comprehension-driven learning loop. When the learner encounters a piece of situated language input, it tries its best to interpret the utterance. Unlike the previous example in Figure 1.1, here the learner does not know most of the function words or any syntax. Naturally, it is unable to produce a very complete interpretation based on the utterance alone, but it makes up for what it lacks in linguistic knowledge with its intention reading abilities and its knowledge about typical events in the world. With each exposure to language input, the learner is able to correlate contextually-obtained information with the sequences of words; each of these acts of hypothesizing a new grammatical construction is referred to in the system as a **composition** operation.

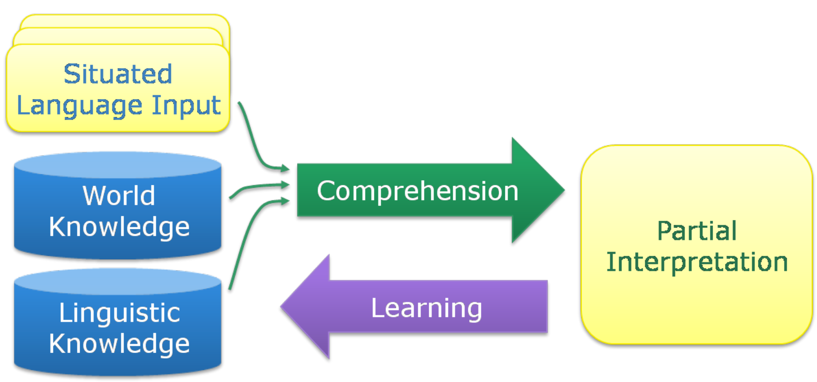

Figure 1.2 A model of comprehension-driven grammar learning. Comprehension of language input is aided by current linguistic knowledge, knowledge about the current situational context, and general knowledge about categories of entities as well as typicality of events.

The contextual information that gets codified into the grammar varies, ranging from role-filler relations between mentioned entities and events to contextual constraints on the usage of the construction. Chapter 2 supplies the details on how contextual information gets turned into new constructions, but the job of the composition operation, at a high-level description, is to create constructions that describe the current situational context as faithfully as possible given the recognized linguistic input. As a result, these initial constructions are very specific to individual lexical items and situational contexts. From these initial constructions the learner is able to build up increasingly more complex constructions and make generalizations, as the next example illustrates. This progression is consistent with Tomasello's item-based construction hypothesis (Tomasello, 2003), which suggests that children initially generalize based on specific words, in particular the main verb, forming constructions such as ___*'s hitting* ___, or *I'm* ___*ing it*, and only later create more abstract constructions by analogy.

While useful as guiding principles for learning, most current theories on construction formation and generalizations, including Tomasello's, underspecify the knowledge and mechanisms required in the process of grammar learning. The ultimate goal of this dissertation is to flesh out these details in precise computational terms. To accomplish this, the model sets out to learn the linguistic conventions of how verb arguments are expressed (i.e. the argument structure constructions) in a language with argument omission, assuming as pre-existing knowledge words for objects and actions, and embodied knowledge about what each action involves. The driving principle in this endeavor is that language is not learned in isolation but in discourse and situational context. A model of situated language learning thus requires the convergence of: (1) a unified representation that captures both semantic knowledge and contextual knowledge, (2) a context-aware language understanding process, and (3) a structured learning and generalization process.

1.1 A preview of the learning model

We give a sketch here about how the learning model works using a few dialogues taken from the BEIJING CHILDES corpus (MacWhinney, 2000; Tardif, 1993; 1996).[3] The example highlights how non-linguistic and linguistic information are combined as learning input to the model, and how constructions learned over time go from specific to general. For simplicity the learned constructions will be given in shorthand notation in the diagram. A technical discussion of how they are represented in the system is offered in Chapter 2.

Dialogue 01 in Figure 1.3 starts out with the example sentence we looked at in Section 1.1. In (a) the father offers the medicine to Xixi while saying *xi1xi1 chi1 yao4* ('xixi eat medicine'). The learner is unable come up with a complete analysis due to its lack of phrasal and clausal grammar knowledge, but is able to recognize the words for the child's name, eating, and medicine and is able to leverage context in putting the meanings together. Judging from intonation that it is requested to perform an action, the learner guesses that it has to do with eating because the word *chi1* ('to eat') is present in the utterance. Medicine is a likely eatee since it is both available in context and mentioned by the father. The learner can then compose a new construction that puts together the three words *xi1xi1*, *chi1*, and *yao4* in that order and denotes an EAT event in which xixi is the eater and medicine is the eatee.

Figure 1.4 puts the composition operation in more concrete terms. Each of the three words are recognized separately due to the lack of phrasal or clausal constructions, and the left half of the figure displays the three corresponding fragments of analysis (solid linked blocks showing the form-meaning mappings). Critically, references to the entities in context are shared between these analysis fragments (indicated by bold dashed arrows from left to right). Namely,

[3] Due to the limited corpus data available, transcripts from several children are combined for use in the learning model.

the word *xi1+x1* refers to the same child in context as the eater of the Eat event denoted by the word chi1, and the word *yao4* refers to the same medicine in context as the eatee of that Eat event.

Dialogue 01

a)	Context:	Father offers child medicine	compose: XI1XI1-CHI1-YAO4
	Input:	xi1+xi1 chi1 yao4 XiXi eat medicine	form: XI1XI1 CHI1 YAO4
	Gloss:	Xixi, take your medicine.	meaning: eater — EAT — eatee

Dialogue 02

b)	Context:	Mother feeds child rice	compose: CHI1
	Input:	chi1 bao3 le ma eat full PRF Q	form: CHI1
	Gloss:	Have you eaten yourself full?	meaning: <Child> — EAT — <Rice>
c)	Context:	Child offers rice to mother	compose: WO3-CHI1
	Input:	wo3 bu4 chi1 1SG NEG eat	form: WO3 CHI1
	Gloss:	I'm not eating.	meaning: eater — EAT — <Rice>
d)			generalize: XI1XI1-CHI1-YAO4, WO3-CHI1 form: {WO3, XI1XI1} CHI1 YAO4 meaning: eater — EAT — eatee form: {WO3, XI1XI1} CHI1 meaning: eater — EAT — <Rice>
e)	Context:	Mother declines rice	compose: NI3-CHI1
	Input:	ni3 chi1 ba 2SG eat SA	form: NI3 CHI1
	Gloss:	Why don't you eat the rice?	meaning: eater — EAT — <Rice>
f)			generalize: XI1XI1-CHI1-YAO4, WO3-CHI1, NI3-CHI1 form: {WO3, XI1XI1, NI3} CHI1 YAO4 meaning: eater — EAT — eatee form: {WO3, XI1XI1, NI3} CHI1 meaning: eater — EAT — <Rice>

Figure 1.3 An example learning sequence that takes place across two dialogues about eating and their corresponding situations starting with only lexical knowledge. The dialogues contain interleaving utterances and events; the sequence of prior events and speech-acts leading up to an utterance provides context for that utterance. With each utterance in context, a new construction is composed if no existing constructions cover the input. Having multiple constructions with comparable meanings in the grammar is a trigger for the generalize operation which creates small grammatical categories (denoted above using curly brackets) that can be used in place of the lexical items.

This sharing of contextual references signifies a meaning relation between previously independent pieces of existing constructions such as XI1XI1 and CHI1, while the utterance, by its sequential nature, supplies the form relations between them. This association of form relations with meaning relations triggers the composition operation from which a new construction, XI1XI1-CHI1-YAO4, is created.

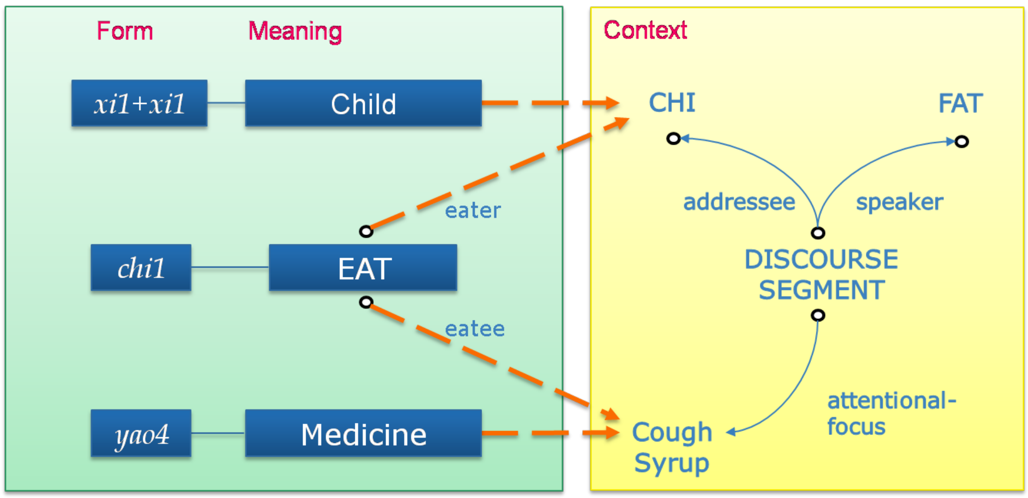

Figure 1.4 The learner's attempt at interpreting *xi1xi chi1 yao4* using only lexical constructions result in an analysis with three fragments or roots. Fortunately, the rich context enables shared references (e.g. to the child and to the cough syrup) between these fragments to be recovered. These shared references form the basis of new compositions.[4]

In (b) in dialogue02, the mother and child are sharing a meal. The mother inquires whether the child is full with the utterance *chi1 bao3 le ma* ('eat full PFV Q'[5]) after feeding him a few spoonfuls of rice. Using the same intentional inference mechanisms but unable (yet) to link up the meaning of EAT and FULL[6], the learner hypothesizes that *chi1* can be used on its own as

[4] A color convention is adopted for figures for those with a color copy of the document. Existing world and linguistic knowledge is depicted with green/blue hues whereas contextually obtained information is depicted in yellow/orange. Learning-related information is in purple.

[5]The gloss generally follows the convention used by Li & Thompson (1981), with a few additions for negation and pronouns. The few glosses used here are: 1SG – first person singular, SA – solicit agreement, NEG – negation, PFV – perfective, Q – question marker.

[6] This is, afterall, an example of a verbal resultative construction which requires some third meaning component such as cause-effect or means-and-ends to explain the semantic relation between eat and full.

long as the eater is a child and the eatee is some rice. These contextual type restrictions on the participants are shown in Figure 1.3 in angle brackets.

The clausal CHI1 construction, despite having only one constituent, is an example of a context-bound construction whose restrictions are relaxed over time through exposure to different usages. It is important to note that the learner, through its experience with the world, understands that there are core participants in events. Eating, for example, involves at least an eater and some food whereas being full involves only one protagonist. In the case of the CHI1 construction when neither role of EAT is filled linguistically, the learner will still try to recover these role fillers from context and remember them as restrictions on the fillers.

The child then offers the rice to the mother, who declines the food by saying *wo3 bu4 chi1* ('1SG NEG eat'). Unsure of how to combine the negation word with the rest of the utterance, the learner can at least put *wo3* and *chi1* together while assuming that the missing core role eatee in the event refers to the rice. This leads to the WO3-CHI1 construction in (c). At this point, an eager learner generalizes based on *xi1+xi1 chi1 yao4* and *wo3 chi1*. By aligning the form and meaning relations, the learner realizes that *xi1xi1* and *wo3* both precede the word *chi1* and are, importantly, connected semantically to the eater role of the EAT event. The learner concludes that either referring expression can be used interchangeably in the XI1XI-CHI1-YAO4 construction and the WO3-CHI1 construction. To do so, a restricted grammatical category consisting of {*xi1xi1*, *wo3*} is created and two new generalized constructions, each with an open slot, are created in (d).

This short example illustrates how information from multiple sources is employed in the learning process. Figure 1.5 depicts how these sources information come together in the implemented system. As depicted in the diagram, both linguistic knowledge and world knowledge are employed in the task of language comprehension performed by the analyzer. Linguistic knowledge takes the form of constructions, which, as mentioned before, are conventionalized pairings of form and meaning that include both the lexicon and phrasal syntax. A specific flavor of construction grammar called Embodied Construction Grammar (ECG) (Bergen & Chang, 2005), which has a precise computational realization and an emphasis on embodied meaning representations, is used by the model and will be explained in the Chapter 2. World knowledge is available primarily through embodied meaning schemas (also specified in ECG) that describe events, frames, relations , spatial configurations, etc, and a corresponding probabilistic model of semantic typicality judgment (e.g. how likely is a ball the throwee of a throw action?). A supplementary ontology is also available to the model to assist inference, and its details will too be described in Chapter 2.

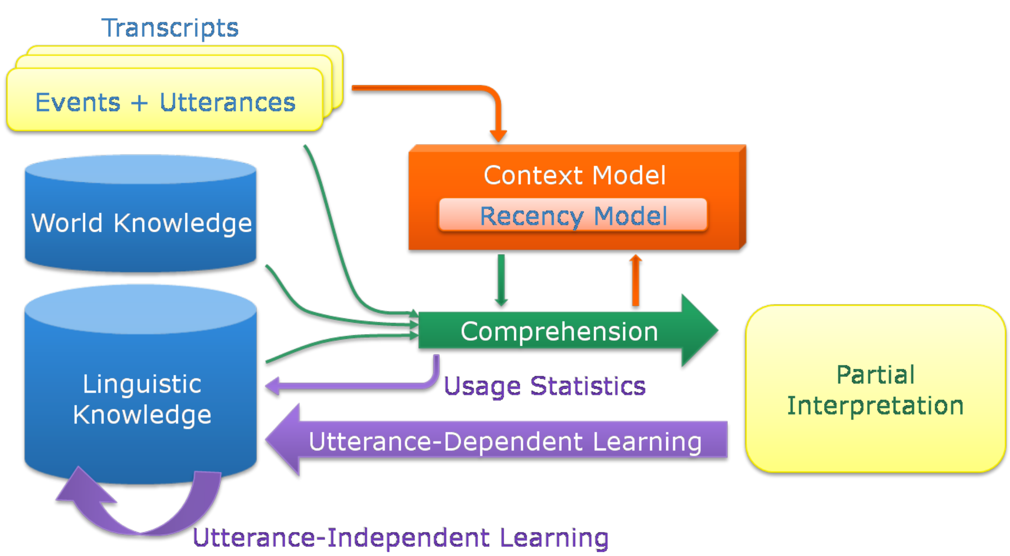

Figure 1.5 **A more fleshed out diagram of the learning model. Utterances in context are analyzed using world knowledge, linguistic knowledge, and support from the context model, and usage statistics are gathered in the process. The output of the comprehension process is a partial interpretation (given that the grammar is incomplete) which is then used to drive a number of utterance-dependent learning mechanisms such as the composition operation. Periodically, the learned constructions are reorganized using other learning mechanisms which are not directly dependent on the immediate input, such as generalization. Although various components of this model are depicted as separate boxes and implemented as separate functions by computational necessity, we believe that a number of these processes occur simultaneously in the brain.**

The combination of linguistic and world knowledge helps the analyzer determine the overall structure of the meaning of an utterance, but an additional context model is needed to relate the interpretation to the current situation. The supplied, dynamically updated, model of context, which keeps tab on events and speech acts that unfolds in a discourse, informs the analyzer of discourse and situationally-relevant entities that may be referred to, particularly in cases of omitted arguments.

The output of this language comprehension process is a partial interpretation containing both a syntactic analysis of the utterance and a meaning representation which is given to the learning processes as input. Learning that happens as a result of this comprehension process, such as composition in the previous example, is labeled Utterance-Dependent Learning. Other learning operations that are less directly tied to an utterance, such as generalization in the last example, are labeled Utterance-Independent Learning operations. The result of these learning operations is new constructions that are added to the repertoire of linguistic knowledge and subsequently used in future iterations of language comprehension.

Leaving the implementation details of these components to Chapter 4 through Chapter 6, I will attempt here an overview of the psychological motivations for the design of this model. From the outset, this work is an attempt to pull together known constraints from different disciplines in outlining a unified view of how language learning may proceed. While a large part of how grammar learning happens remains to be discovered, a fair amount is known about the cognitive development of a typical child up to the language learning age. These findings shed light on the contribution that a child brings to the task of learning language. The findings most relevant to this thesis are in the realms of social abilities, semantic knowledge, and learning mechanisms that are in a sense prerequisites for grammar learning.

1.2 Developmental support for grammar learning

The grammar learning process in the model draws on a number of domain-general abilities that children have by the time they start positing syntactic and semantic relations between words. These are:

1. Formation of object and event categories
2. Understanding of other people's goals and intentions
3. Rule learning, fast mapping, and structure mapping
4. Statistical learning mechanisms

In particular, children's early words and concepts form the basis of the model's initial knowledge. Children's ability to understand other people's goals and intentions lends support to the model's use of rich contextual information in figuring out the intended meaning of a caregiver's utterance. The model's construction formation and generalization capabilities are supported by evidence of children's ability to create rules, map structurally across form and meaning domains, and analogize between constructions through structural mapping. Finally, the model is able to learn probabilistic preferences on constructions just as children are sensitive to statistical regularities in their input.

Formation of object and event categories

Categorization of objects appears rather early in development. Infants come to a rudimentary understanding of objects and their physical properties starting at 3.5 months, appreciating the principle of continuity, object permanence, etc (Baillargeon, 2004; Spelke, 1994; Munakata et al, 1997). In learning object labels and forming categories, 16- to 21-month old infants attend selectively to cues that are informative of the category boundaries, shape being one of these cues (Welder & Graham, 2001). How infants conceptualize events is currently a bit less well understood. Mandler (1992) outlines a set of conceptual primitives based on image schemas (Johnson, 1987) which includes containment and support, and intersects notions of self motion and caused motion with animacy and agency. (Golinkoff, Hirsh-Pasek, Mervis, Frawley & Parillo, 1995) reviews the experimental evidence for the development of these conceptual primitives and offers an account of how these primitive notions of events can be recruited in verb learning.

Both object and event representations get more fine-grained as children grow older (Munakata, McClelland, Johnson & Siegler, 1997; Wilcox & Schweinle, 2002), and so do the capacity for mapping verbs to events. For example, the understanding of causality emerges at around age 2, but it is not until around age 5 that children have a good grasp of the causal relationships in everyday physics (Gopnik, Glymour, Sobel, Schulz, Kushnir & Danks, 2004). In an experimental setting, 18-month-olds were found to require explicit cues in order to extract the relevant dimensions of a labeled action and use them to extend the label to another actor performing the same action (Maguire, Hennon, Hirsh-Pasek, Golinkoff, Slutzky & Sootsman, 2001). In a separate experiment, 22-month-old infants were unable to override perceptual cues and learn a new action label based on sociolinguistic cues unlike the 34-month-olds subjects (Brandone, Pence, Golinkoff & Hirsh-Pasek, 2007).

In light of these findings, the current model does make some aggressive assumptions about the structure and organization of the linguistically-relevant semantic knowledge available to the learner at the start. However, since the model is also focused on later, more sophisticated linguistic phenomena (e.g. coverb-marked argument displacement), the assumed knowledge is in line with the studies of the slightly older toddlers.

Understanding of other people's goals and intentions

Knowing that parental utterances are intended for communication and being able to decipher the intended meaning of an utterance given context is fundamental to the model's ability to learn grammar. Afterall, it is the gap in understanding between what the model can derive from its current grammar and what it thinks the utterance is intended to convey that drives the hypothesis of new constructions. Support for this approach can be found in children's fairly robust social ability by the time they are one year of age. Infants as young as 3 months old (Hood, Willen & Driver, 1998) can reliably follow the eye gaze of an adult and infer the goals of other people's actions by 6 months. They are capable of holding joint attention on an object with a caretaker by 9 months, which coincides with the age infants start to produce their first words. Tomasello argues that the ability to conceive of the triadic relation between self, another person and a third object, as well as the ability to think of other people as intentional agent like the self, is what allows infants to understand language as a communicative act (Tomasello, 1999; 2001).

There is ample evidence that children do not rely blindly on learning by association when it comes to language. Children do not just associate pieces of language input with whatever they happen to be attending to but instead care about the goals of the speaker (Bloom, 2002). When an experimenter tries to illustrate a novel word with a particular action, fails to complete the action and expresses surprise, children do not mistake the failed action for what the word is intended to label (Akhtar & Tomasello, 1996). The learning model in particular relies on an intentional zoom-lens when it comes to postulating the meaning of a new construction: there is simply too much going in any given scene (even given all the computational simplifications) and relevant contextual information must be extracted in a sensible way.

Rule learning, fast mapping, and structure mapping

In the learning model, the composition of a new construction is essentially learning a new rule, albeit one that involves structured constraints in two domains. One of the first investigations of infants' rule learning ability was done by Marcus and his colleagues, where they found that 7-month-old infants were able to abstract patterns of the form ABA, ABB or AAB from exposure to streams of syllable sounds and discriminate between novel stimuli of the familiar pattern from an unfamiliar one (Marcus, Vijayan, Bandi Rao & Vishton, 1999). This rule-learning ability was recently found to be not limited to linguistic input: 12-month-old infants were found to able to learn rules governing picture sequences of dogs and cats (Saffran, Pollak, Seibel & Shkolnik, 2007).

Fast mapping is another domain-general learning mechanism that has been shown to be useful to word learning (Heibeck, 1985; Markson & Bloom, 1997). It is the idea that a few incidental exposures are enough for the learner to form a long-lasting association between two stimuli (e.g. an object and a label). Fast mapping on a phrasal level is only beginning to be investigated for 5- to 7-year-old children (Casenhiser & Goldberg, 2005).

Finally, analogy and structure mapping have been proposed by a number of researchers as the primary means of creating generalization in a grammar (Gentner & Markman, 1997; Gentner & Namy, 2006; Macwhinney, 2004; Tomasello, 2000). Specifically, Tomasello (2000) proposes that a child learns verb island constructions through structure mapping between form and meaning, and through further processes of structure mapping creates more and more abstract constructions (e.g. the simple transitive constructions and then the subject-predicate construction). However, having noted that both form and function are critical to structure mapping, he conceded that it is still not known what a "good" structure map between construction entails, and whether some "critical mass" of exemplars are required before abstraction through structure mapping can take place.

Statistical learning mechanisms

Another line of research that has played an important role in language acquisition is that of statistical learning. Whereas rule learning is concerned with how children create categories of items (also sometimes called algebraic rules), statistical learning is generally concerned with the how children track frequencies and probabilities associated with these items (e.g. syllables and words).

Even before they reach 1 year of age, infants are able to extract various statistical regularities in their linguistic as well as visual input (Aslin, Saffran & Newport, 1998; Fiser & Aslin, 2002; Kirkham, Slemmer & Johnson, 2002; Saffran, Aslin & Newport, 1996). Their ability to track these statistical regularities multiple domains suggests that statistical learning is a powerful domain-general learning mechanism that gives infants a head start in learning language.

In particular, infants' ability to make use of transitional probabilities between syllables and phonological patterns is argued to help infants segment words from their input (Chambers, Onishi & Fisher, 2003; Maye, Werker & Gerken, 2002; Saffran & Thiessen, 2003). There is also evidence that the output of this kind of statistical learning are word-like representations that are readily integrated into the native language (Saffran, 2001) and that structures like those found in natural languages are readily acquired by 12-month-old infants (Saffran, Hauser, Seibel, Kapfhamer, Tsao & Cushman, 2008).

Studies within the statistical learning paradigm have been extended to investigate whether children can also extract statistical regularities in non-adjacent dependencies, which are seen as the basis of acquiring certain properties of phonology (such as vowel harmony) as well as syntax (such as agreement) (Gomez, 2002; Newport & Aslin, 2004). Gomez found that given a pre-segmented speech stream, both adults and 18-month-old infants are able to learn the dependencies between non-adjacent words if there is a high variation of the intervening words. More recently, Thompson and Newport (2007) studied the roles of different structural cues in the statistical learning of phrase structure. In particular, they found that structural variations such as optional phrases, repetition, movement, class size (when used alone and more so in combination) created the right kind of transitional probability variations at phrase boundaries that allowed their adult subjects to successfully learn the syntax of a fairly complex artificial language.

Wonnacott, Newport & Tanenhaus (2008) examined the role of distributional learning in acquiring argument structure alternations and their subcategorization restrictions. To use their example, the verb throw can participate in both the caused motion construction and the ditransitive construction, as in *Jack threw Henry the ball* and *Jack threw the ball to Henry*, but semantically similar verbs such as *carry* and *push* cannot participate in the ditransitive construction, as in **Jack carried/pushed Henry the ball*. They found that in the absence of consequential semantic distinctions, adult subjects are still able to generalize across the stimuli and learn both absolute and probabilistic subcategorization constraints.

With this, we are almost ready to begin the exposition of the learning model. But first, we will highlight some important properties of Mandarin Chinese that make it an interesting object of study for this work.

1.3 Case study: Mandarin Chinese as a pro-drop language

Although the current learning model makes few assumptions that restrict its cross-linguistic application, Mandarin Chinese is used as the primary language in testing the model for its context-dependent properties and relative simplicity in morphology. Like English, Chinese is a predominantly Subject-Verb-Object (SVO) language. However, Chinese differs from English in several notable ways that have implications for any computational system that purports to do language understanding and language learning. This section gives an overview of aspects of Mandarin Chinese relevant to the learning model. The interested reader can refer to Erbaugh (1992), Lee (1996), and Li & Thompson (1981) for more in-depth background of Chinese linguistics and the acquisition of Mandarin Chinese.

Subject-drop and object-drop

Unlike English, subject-dropping and object-dropping is freely allowed in Chinese. Wang, Lillo-Martin, Best & Levitt (1992) recorded conversations between five native adult Mandarin speakers with another adult and found that 45.6% of sentences have an omitted subject and 40.1% of them have an omitted object. (There is unfortunately no data available on what percent of these sentences have both omitted subjects and objects.)

The subject-drop phenomenon in Chinese is often compared to that in Spanish, Italian, or Japanese. However, unlike these other languages, Chinese does not require verb agreement or case marking, making the disambiguation of the referent of the omitted subject heavily dependent on the discourse and situational context. Kim (2000) compiled a cross-linguistic comparison of subject omission, adapted here:

English	Portuguese	Italian	Mandarin	Cantonese	Korean	Japanese
2 – 4%	44%	44 – 54%	36%	42%	55 – 65%	62 – 74%

Figure 1.6 Percentage of adult utterances with omitted subjects in seven languages, adapted from Kim (2000).

To compare the adult argument omission pattern to child-intended speech, in the same study Wang et al also asked adult subjects to narrate a story book while pretending that they are speaking to their own child. The rate of subject omission was measured to be 36.13% and object omission, 10.3%.

Children's sensitivities to the adult patterns of argument omission are reflected in their production from a young age. Even though children in all languages start out producing fewer subjects than adults do, even in a non-pro-drop language like English, English-speaking children soon start producing more subjects whereas pro-drop language learners start matching the omission rate of the adult speech. In particular, 2- to 2.5-year-old Mandarin-speaking children omit subjects 56% of the time, 3- to 3.5-year-olds omit subjects 46% of the time, and those who reach 4 years old omit subjects only 38% of the time.

Topic-comment and topic chain

Chinese is argued to be a topic-comment language, and notions of topic and topic-chains have been proposed (Li & Thompson, 1981; Li, 2004; Shi, 2000). The topic of a sentence, roughly speaking, is what the sentence is about, and it either has been introduced in the discourse (definite)[7] or is a generic. The topic need not coincide with the subject of a sentence and are sometimes marked by topic markers, such as *a*, *me*, and *ne*. Here is an example:

nei4	*ben3*	*shu1*	*a,*	*wo3*	*yi3+jing*	*kan4*	*wan2*	*le.*
that	CLS	book	*a*	1PS	already	see	$\text{ASP}_{\text{finish}}$	CRS.

That book, I have already finished reading.

The notion of topic becomes quite important when resolving the intended references of omitted subjects or objects. The idea of a topic chain, as proposed by Li & Thompson (1981), is

[7] The original formulation in Li & Thompson (1981) does not include situational context, but the data from the child language corpus shows that a salient person/object/event in the situational context can serve as the topic as well.

that a referent (the topic) is introduced in the first clause, and subsequent clauses can refer to the topic without explicit mentions. With evidence from written narrative text, W. Li (2004) proposes a modified notion of topic chain where the topic of a chain need not even appear in the initial clause. The topic can be coreferential either anaphorically with a previous sentence, or cataphorically with a noun phrase later in the sentence.

Coverbs

Chinese uses coverbs, a closed class of grammatical words, to express the equivalent of case relations, such as object, dative, benefactive, locative, ablative, and terminals (Erbaugh, 1992; Li & Thompson, 1981). Coverbs precede a noun phrase and the resulting coverb phrase generally precedes the main verb, though iconicity sometimes influences the word order. As an example, whereas *cong2*, which marks the origin, comes before the main verb, *dao4*, which marks the destination, comes after the verb, in accordance with the iconicity of motion. This need not be the case, as exemplified by *wang3*, which marks the direction and comes before the verb. Here are examples of these three coverbs being used:

ba3	*qiu2*	*cong2*	*na4*	*bian1*	*na2*	*guo4*	*lai2.*
CV_{obj}	ball	CV_{origin}	that	side	carry	across	towards.

Bring the ball over here from there.

ba3	*qiu2*	*fang4*	*dao4*	*na4*	*bian1.*
CV_{obj}	ball	put	$CV_{destination}$	that	side

Put the ball there.

ba3	*qiu2*	*wang3*	*na4*	*bian1*	*reng4.*
CV_{obj}	ball	$CV_{direction}$	that	side	throw.

Throw the ball towards there.

Coverbs are historically derived from full, free verbs. A handful of coverbs, such as *ba3* (direct object marker) or *bei4* (agent marker), can no longer be interpreted as full verbs, but other ones can take aspect markers even when functioning as case markers. Some of these coverbs retain a separate present-day sense as full verbs, such as *gei3* (benefactive / recipient marker) which can also mean to give. This confusion in coverbs is problematic for children acquiring Mandarin, and the bleached coverbs such as *ba3* are sometimes mistakenly used as full verbs (Erbaugh, 1992).

The prevalence of argument omission together with the use of topic-comment constructions as well as a coverb that allow objects to be fronted suggests that word order is not as reliable a cue of semantic relations as it may be in other word order languages.

Serial verbs and conditionals

The serial verb construction is often used in Chinese, where two or more verb phrases or clauses are conjoined without any marker indicating what the relation is. The form of the construction looks like (parenthesis denoting optional elements):

(NP) VP (NP) (NP) VP (NP)

Li & Thompson (1981) categorizes the serial verb construction into four groups:

i. Two or more separate events, which can be related as a temporal sequence, a causal sequence, co-occurrence, or the circumstances of occurrence.

(NP) VP (NP) (NP) VP (NP)

ii. One verb phrase serving as the subject or direct object of another verb, similar to sentential complements in English.

(NP) VP (NP) VP (NP)

direct-object

(NP) VP (NP) VP (NP)

subject

iii. Pivotal constructions, where the intervening noun phrase serves as both the direct object of the first verb and as the subject of the second verb.

(NP) VP NP VP (NP)

direct-object subject

iv. Descriptive clauses, where the second verb phrase describes or modifies the direct object of the first verb phrase in some way.

(NP) VP NP (NP) VP (NP)

describes

In addition, a further potential point of confusion is that the Mandarin conditionals overlap strongly with the serial verb construction. In Mandarin conditionals, there are no verb tense or mood changes, and the two clauses (the premise and the conclusion) can be expressed without any conjunctions or conditional markets. Therefore, analogous to noun-noun compounds in English, deducing the relation between two conjoining verb clauses in Mandarin Chinese can be a difficult inference problem. We will see that the model has difficulty with this as well. In the remainder of this section, we summarize the developmental trajectory of children acquiring Mandarin Chinese, drawing on surveys done by Erbaugh (1992) and Lee (1996).

Acquisition of Mandarin Chinese

Erbaugh (1992) based her studies on 64 hours of longitudinal data with four Chinese children (1;10 – 3;10) in Taiwan and concluded that children generally followed the canonical SVO word order throughout development and were reluctant to violate the SVO order, especially when they were young, even in situations where an OV re-ordering (using the *ba3* direct object marker) is required. A summary of her findings is presented in Figure 1.7. Lee (1996), on the other hand, argued that Mandarin-speaking children's word order was not as rigid as Erbaugh suggested, citing other studies that showed children producing incorrect OV sentences such as *Di4+di ma4* ('brother scold') (2;0) in answering "who did Ah San scold?".

Age	Developmental Characteristics
before 2	▪ Mean length of utterance (MLU) < 2.0 ▪ nouns: kinship terms, concrete nouns ▪ verbs: actions, activities, statives ▪ rare use of modals
around 2	▪ 1.8 < MLU < 2.5 ▪ strong SVO order: producing either SV or VO chunks ▪ rare use of the complete SVO ▪ topicalization to OV not present ▪ some uses of SV to indicate patient state ▪ use of the modal *yao4* (want)
2;3 to 3;2	▪ 3.0 < MLU < 4.0 ▪ strong SVO order ▪ topicalization of direct object with *ba3*, and agent with *bei4* ▪ serial verbs: *lai2* ('come'), *qu4* ('go'), *gan3* ('dare'), *bang* ('help') ▪ benefactive *gei3* appears ▪ use of some modals: *hui4* ('can', 'might') and *neng2* ('can', 'able to') ▪ aspect: perfective *le* acquired before durative *zai4* and progressive *zhe* ▪ event time rarely marked ▪ duration and manner rarely present ▪ mistakes particles/coverbs for full verbs ▪ overuses the general *ge4* classifier ▪ special classifiers associated with prototypes ▪ yes/no questions of the form A-NOT-A appears (using modals) ▪ wh-questions first appear with *shen2me* ('what')
after 3;2	▪ MLU > 4.0 ▪ full sentence syntax ▪ modals and serial verbs fully present ▪ aspect: use of past experiential marker *guo4*, progressive still difficult ▪ event ordering is implicitly present in verb ordering ▪ more special classifiers emerge, though problematic till as old as 7 ▪ some use of discourse-sensitive particles, topicalization, and re-ordering.

Figure 1.7 Summary of the linguistic development of Mandarin-speaking children (adapted from Erbaugh, 1992)

As an additional confound to the acquisition of word order, it is unclear from the two studies how frequently children confused word classes. Erbaugh reported that her subjects never used nouns as verbs, although they sometimes confused adjectives or adverbs with verbs. In citing errors made by the child Laohu (2;0), who used *wo3 ji1qi4ren2 le* ('I robot PFV') to mean “I have become a robot”, she suggested that the child had incorrectly suffixed the perfective marker *le* to nouns and omitted the main verb. However, Lee suggested that a different method of error accounting can attribute the error to the child's confusion between nouns and verbs, and showed examples of other mistakes of this kind, such as a child (2;4) using *bey shou3 deng1* ('baby hand lamp') to mean “baby touched the lamp”.

This brings up another point of contention in how Mandarin Chinese is acquired — whether verbs are really verbs (Bates, Chen, Tzeng, Li & Opie, 1991; Li, Jin & Tan, 2004). This question arose in the context of an apparent lack of a noun bias in Chinese(Tardif, 1996; Tardif, Shatz & Naigles, 1997), as reflected in the much higher proportion of verbs in the early vocabulary of Mandarin-speaking children. Tardif (2006) argued that early verbs in Chinese are indeed verbs by virtual of distinct syntactic markings used by children on the verbs (e.g. negation, resultative verb complement). Instead, Chinese verbs are easier to acquire because they are more imageable (Ma, Golinkoff, Hirsh-Pasek, McDonough & Tardif, 2008), and this is indirectly confirmed by a replication of the human simulation experiment (Gillette, Gleitman, Gleitman & Lederer, 1999) in Mandarin (Snedeker, Li & Yuan, 2003). In these experiments, adult subjects see videos of caregiver interaction with the audio removed and have to guess the word that the caregiver is using wherever a beep is signaled. English-speaking subjects were much worse at guessing English verbs than English nouns (Gillette et al., 1999), underscoring the difficulty in inferring the intended meaning of verbs. However, when asked to perform the same task on muted videos of Mandarin-speaking mothers playing with their children, both English-speaking and Mandarin-speaking subjects were just as good at guessing the muted verbs as the nouns.

1.4 Summary

What is beginning to emerge is a very complicated picture of how a language is learned by children, especially when cross-linguistic variations are taken into account. In the current modeling endeavor we will hold fast to principles of domain-general learning and attempt to show how one small piece of languge learning — argument structure constructions in Mandarin Chinese — can be learned. Chapter 2 lays out the representational foundation for this work and show how situated language understanding can be modeled given this representation. Chapter 3 gives the precise computational definition of the learning problem. The details of how the learning model implements each learning operation and updates its statistics on the grammar are given in Chapter 4 through Chapter 6. Experimental results on a subset of the Beijing CHILDES corpus of parent-child interactions are reported in Chapter 7 and follow-up experiments using artificial languages are reported in Chapter 8. A general discussion of the learning model and ideas for future work are offered in Chapter 9.

Chapter 2.

Understanding an Utterance in Context

The last chapter introduced the basic formulation of the comprehension-driven learning system: how semantic structures obtained from context are leveraged to create syntactic and semantic structures (constructions) in the grammar, and how these constructions are subsequently generalized. This chapter lays out the foundations for describing the learning model in technical detail. The model is an integral part of the NTL simulation-based language understanding paradigm (Chang, Feldman & Narayanan, 2004; Feldman, 2006; Narayanan, 1999) which stipulates that language understanding requires active simulation of the meaning representation in addition to the constructional analysis of a sentence. For example, the constructional analysis of

Put the bear in the basket.

creates a specification of thematic role-filler relations (e.g. the addressee is the putter, and the bear is the puttee, and the basket is the goal). Simulation, which is a process of active inference, generates additional predictions about actions, causes, and consequences (e.g. the bear is likely to be contained by the basket at the end of the put action and will be transported along with the basket if the basket is moved, unless the basket is not big enough for the bear). The outcome of simulation is therefore a product of linguistically-derived cues (such as the thematic role-filler relations) and ontological knowledge (such as knowledge about the size of a typical toy bear or a specific bear in context).

The embodied simulation-based language understanding hypothesis posits that active simulation is carried out using the same neural substrates that underlie motor actions. In computational terms, a mental representation of the state of the world is updated continuously through monitoring of the situational context as well as eager integration of constraints from linguistic input. These convergent constraints often reduce ambiguity in the input and help a language user arrive at what seems to be the obvious interpretation. The dynamically updated embodied mental representation also allows for additional inference about the state of the world as it existed in another time, place, or mind.

The empirical support for online integration of constraints and online update of mental representation comes from a number of psycholinguistic experiments not traditionally associated with the embodied simulation view. A bulk of this work is performed in the visual world paradigm using eye-tracking techniques, where a static picture depicting various people and/or objects is presented to the subject as an auditory sentence unfolds. The subject's eye movements to various objects are taken to be reflective of the subject's processing vis-à-vis visual attention. Adult subjects have been demonstrated to be sensitive to contextual cues such as affordances of surrounding objects, taking into account sizes of depicted containers in a visual scene (Chambers, Tanenhaus, Eberhard, Filip & Carlson, 2002) when movement of an object is described, or whether a glass contains remaining wine when a drinking event is described in the past tense (Altmann & Kamide, 2007). Similar results on the effect of visual context have also been obtained in the mouse movement paradigm where subjects are asked to move objects on a computer screen based on an auditory stimuli (Spivey, 2008; Spivey & Dale, 2006). Altmann has additionally shown that subjects make note of and look to linguistically updated locations of depicted objects even though the visual scene remains static, indicating an integration of linguistic and visual information (Altmann & Kamide, (under review)).

The empirical evidence for embodied representations supporting the active simulation comes from a host of studies of the mirror neuron system. These studies suggest shared mechanisms for both action observation and action execution (Gallese, Fadiga, Fogassi & Rizzolatti, 1996; Murata, Fadiga, Fogassi, Gallese, Raos & Rizzolatti, 1997), and there is growing evidence that the same system is also utilized in language understanding (Buccino, Riggio, Melli, Binkofski, Gallese & Rizzolatti, 2005; Gallese & Lakoff, 2005; Skipper, Goldin-Meadow, Nusbaum & Small, 2007). In behavioral studies, Bergen found a facilitation effect for directional motor movements congruent with the direction specified by linguistic stimuli and interference for incongruent stimuli (Bergen, Chang & Narayan, 2004; Bergen & Wheeler, 2005), again suggesting that action language may (partially) activate the neural circuitry used to execute actions. Matlock's work on understanding fictive motion sentences (e.g. *the path runs along the river*) indicates that processing times of these sentences correlate with the difficulty of the terrain of the path imagined, further suggesting that even metaphorical usages of action words are understood in terms of their grounded, physical meaning (Bergen, Lindsay, Matlock & Narayanan, 2007; Matlock, 2004).

The embodied simulation hypothesis and the associated language understanding framework have important implications for the learning model. A crucial step in the comprehension-based learning loop is extracting as much of a coherent interpretation as possible from the input given an impoverished grammar, and as such the model makes use of a best-fit constructional analyzer (Bryant, 2008a) (properties of which will be described in further detail in the next sections). Both the constructional analyzer and the learning model represent grammar in the Embodied Construction Grammar (ECG) formalism (Bergen & Chang, 2005), which will be described briefly in the next section along with extensions to the formalism that provide a tighter link to context. Readers who are interested in the linguistic details of the ECG formalism can refer to Feldman, Dodge & Bryant (to appear).

Unfortunately, due to implementation constraints in the current system, the constructional analysis process and the simulation/context-based inference procedure are implemented as sequential processes. This simplification will not detract from the main findings of the learning model, however. A fully parallel model of language understanding will outperform the current model in its ability to extract coherent interpretations from context, and learning results are expected to improve given a better language comprehension mechanism.

2.1 Representing Context in Embodied Construction Grammar

Central to the simulation-based understanding paradigm is a theory of grammar that puts syntax and semantics on equal footing and allows constraints from both domains to be incorporated simultaneously during sentence processing. Construction grammar, proposed first by Kay and Filmore and further developed by Goldberg (Goldberg, 1995; Kay & Fillmore, 1999), does exactly that. The basic unit of grammar according to the construction grammar view is a construction — a conventionalized pairing of form and meaning. Constructions encompass not only phrases and clauses but also lexical items, and can range from fixed idioms (e.g. *kick the bucket*) to semi-productive constructions (e.g. *What's X doing Y*[8] or the *way* construction[9]) to

[8] The WXDY construction [Kay Filmore] expresses both surprise and disapproval, as in *what's a nice girl like you doing in a place like this?* or *waiter, waiter, what's the fly doing in my soup?*

[9] The way construction uses a path phrase in conjunction with a manner of motion or activity verb to express the means of achieving a goal, as in *he waltzed his way through the party* or *she talked her way out of the difficult situation*.

fully productive argument structure constructions (e.g. the ditransitive or double-object construction[10]). Another central feature of construction grammar is its allowance for non-compositional meanings. Non-compositional meaning refers to meaning attributed to a construction that cannot be further decomposed into its constituents, making the whole greater than the sum of its parts. Unlike other constraint-based grammars such as LFG, whose meaning-carrying units are lexical items only, construction grammar allows both compositional meaning and non-compositional meanings to be expressed at all levels of constructions. The ability to attach meanings to all levels of constructions is essential for capturing linguistic facts such as the surprise and disapproval in the WXDY construction or the transfer scene in the ditransitive construction.

As a unification-based construction grammar formalism, ECG is not only capable of representing form-meaning mappings but also grounds the meaning representations in embodied schemas. Language understanding and reasoning are thus performed in terms of basic motor schemas, image schemas, and frames, whereas abstract reasoning is enabled by metaphors (Lakoff, 1987; Lakoff & Johnson, 1980; Langacker, 1990; Talmy, 2000). It is important to stress that these schemas, assumed to have at least partially developed in children when they begin to use language, provide the substrate for the learner to both understand the unfolding events and to create mappings from form to meaning.

ECG basic: schemas, constructions and SemSpecs

The four primitive types in ECG are: schemas, constructions, maps, and situations (see Appendix A for the technical specification of the formalism). Of these, **schema**s (the basic unit of

[10] The English ditransitive construction expresses a meaning of transfer or intended transfer which is supplied not by any lexical items but by the construction itself, as in *I wrote him an email* or *she baked me a cake*. There are, of course, semantic restrictions on what verbs can be used in the construction, e.g. **Mow me a lawn.* [Goldberg?]

meaning) and **constructions** (the pairing of form and meaning) are directly relevant to this research and are described in this section. Schemas can be divided into two distinct and important kinds: conceptual schemas and structural schemas. Conceptual schemas are the bulk of the embodied knowledge, such as the aforementioned motor schemas, image schemas, and frames. Structural schemas, on the other hand, are computational conventions for the bookkeeping of information that is important to the constructional analysis process. These include the EVENT_DESCRIPTOR schema and the referent descriptor schema (RD) which will be introduced later in this section. Constructions can be lexical or phrasal/ clausal (i.e. they have constituents) and their meanings are defined in terms of the set of conceptual schemas.

Additionally, an ontology defined externally to the grammar is also provided to the learner (the detailed implementation of the ontology is not germane to this dissertation; its full specification can be found on the ECG wiki at http://ecgweb.pbwiki.com). Whereas ECG schemas capture aspects of sensorimotor and other knowledge that are linguistically relevant, the ontology captures general knowledge that may be relevant for simulation and can be accessed by the grammar if necessary (using the special @ symbol; see further illustrations later in this section). We may represent the toy bear in our first example as a BEAR schema which states its status as a physical object (perhaps a small, movable physical object) but we may also like to store other facts about the specific bear in question, e.g. it was a gift from grandma for the child's last birthday. These additional facts can be stored in the ontology.

During analysis of a sentence, a semantic specification (or SemSpec) is created as its meaning representation. The SemSpec is a network of instantiated schemas and ontology items created from the meaning poles of the recognized constructions, with the appropriate roles from various schemas unified and filled in. These schematic representations of events can in turn be used for simulation. Simulation is computationally realized using X-nets, which are Petri-net based active structures for event-based asynchronous control that can capture both sequential flow and concurrency (Narayanan, 1997). Simulation is a dynamic process which includes executing the X-nets specified in the SemSpec and propagating belief updates in a belief network. Returning to the *put the bear in the basket* example, a put action involves a putter, a puttee, a trajectory of the puttee and a manner of motion. These are specifiable by language and are represented in an ECG schema. However, lower-level motor control parameters of a put action, including the arm movements, hand shape, and the weight of the object, can be captured in an X-net. Work within the NTL group illustrated correspondence between hierarchies of motor action and argument structure constructions (Feldman et al., to appear) and provided understanding of metaphoric language through mappings on embodied simulation (Narayanan, 1999). This thesis will thus take as given low-level representations of actions (Bailey et al., 1997) and focus on learning more complex linguistic structures using relatively high-level, schematic descriptions.

As mentioned in the last section, multiple facets of constructional analysis occur simultaneously in the ideal system: (1) recognizing instances of constructions in an utterance and putting together the meaning schemas according to their unification constraints (analysis), (2) resolving references (resolution), (3) filling in additional details about events based on knowledge about the context (context fitting), and (4) making active inferences to resolve ambiguities and to update the model of context (simulation). As implemented, these processes are carried out in three sequential stages: (i) analysis and resolution in parallel, preserving ambiguities as separate analyses, (ii) context fitting to find the analysis that best matches the context (which also forms the basis for learning), and (iii) simulation.

The remainder of the section demonstrates this framework of understanding language in context using an example in Mandarin taken from the CHILDES Beijing corpus (MacWhinney, 2000; Tardif, 1993; Tardif et al., 1997). The mother directs her child to give a peach to the investigator, who is referred to as *yi2* (aunt):

ni2 gei3 yi2 (2PS give aunt).

Figure 2.1illustrates the desired analysis. The top half shows the constructional tree and the bottom half shows a feature structure representation of the semantics. The sentence is recognized as an instance of the ACTIVE_SUBJECT_VP construction whose subject is filled by the construction NI3 and the verb phrase is filled by a DITRANSITIVE_VP, which takes one verb and two noun phrases as constituents. The main verb in the DITRANSITIVE_VP is filled by the verb GEI3, the first object NP by YI2 and the second object NP is unfilled.

The overall meaning of the sentence is an instance of the EVENT_DESCRIPTOR schema named EVENT_DESCRIPTOR01, which describes an event as the name suggests. Its profiled_process, given by the verb phrase, is a giving action captured by the schema instance GIVE02 (bracketed numbers denote co-indexation). The giver, recipient, and theme roles of GIVE02 are of types @Human, @Aunt, and @Entity respectively. The RDs (RD03, RD04, and RD05) are helper schemas that direct the analyzer to look up the three referenced entities in context; how this is accomplished is described in the rest of this chapter. As a convention, schema and constructions are written in all-caps in this dissertation.

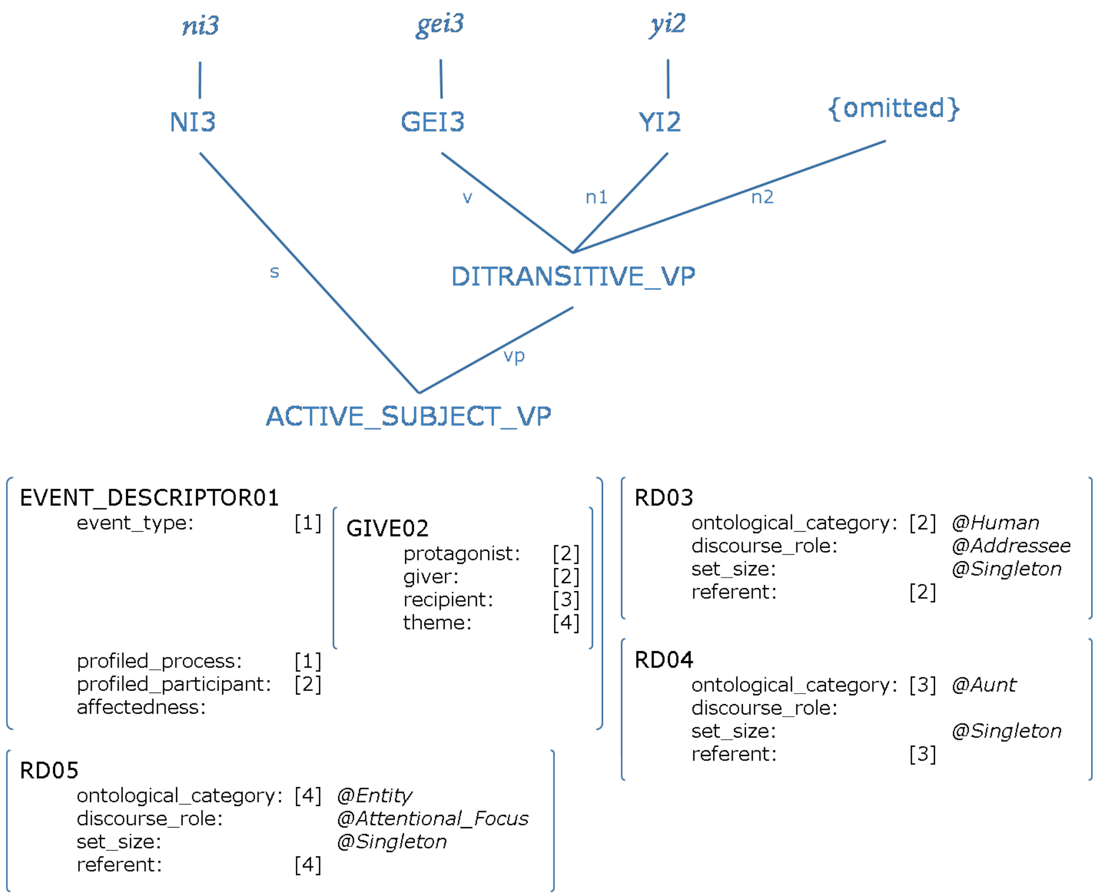

Figure 2.1 The constructional tree and semantic specification (SemSpec) for the ideal analysis of *ni3 gei3 yi2* (you give aunt). The constructional tree shows the use of the DITRANSITIVE_VP construction as the vp constituent of the ACTIVE_SUBJECT_VP construction. The corresponding meaning, created jointly with the constructional tree, can be described using feature structure notation, as shown in the bottom half.

Feature structure notation

The feature structure notation in the bottom half of Figure 2.1 is a standard Computer Science technique of representing structured information[11]. Basic familiarity with feature structures and unification grammar is assumed for the dissertation, but terminology will be established here for clarity. Each schema has roles (also referred to as features). A role establishes

[11] The reader can also refer to [Bryant, 2008] for additional technical details on how this feature structure representation is used in constructional analysis.

a pointer to a slot (or a placeholder) that may have a type constraint (indicated in italics) as well as an atomic or structured value (indicated through coindexation using the bracketed numbers).

Each feature structure can be visualized as a directed acyclic graph (DAG), as shown in Figure 2.2. Outgoing edges indicate features and nodes indicate slots, and conveniently, multiple incoming edges into a node signify coindexation. In this DAG representation, slot chains can be easily understood as following edges through the graph to a particular slot. For example, the 5-way coindexation between GIVE02.protagonist, GIVE02.giver, RD03.referent, RD03.ontological_category and EVENT_DESCRIPTOR01.profiled_participant are shown as the edges pointing to an unfilled slot whose type constraint (in italics in Figure 2.2) is @Human. The identity of the human is not known and therefore the slot is not filled with a value. At the end of the analysis and context-fitting process described in this chapter, the contextual filler of this slot is identified as the child in the scene.

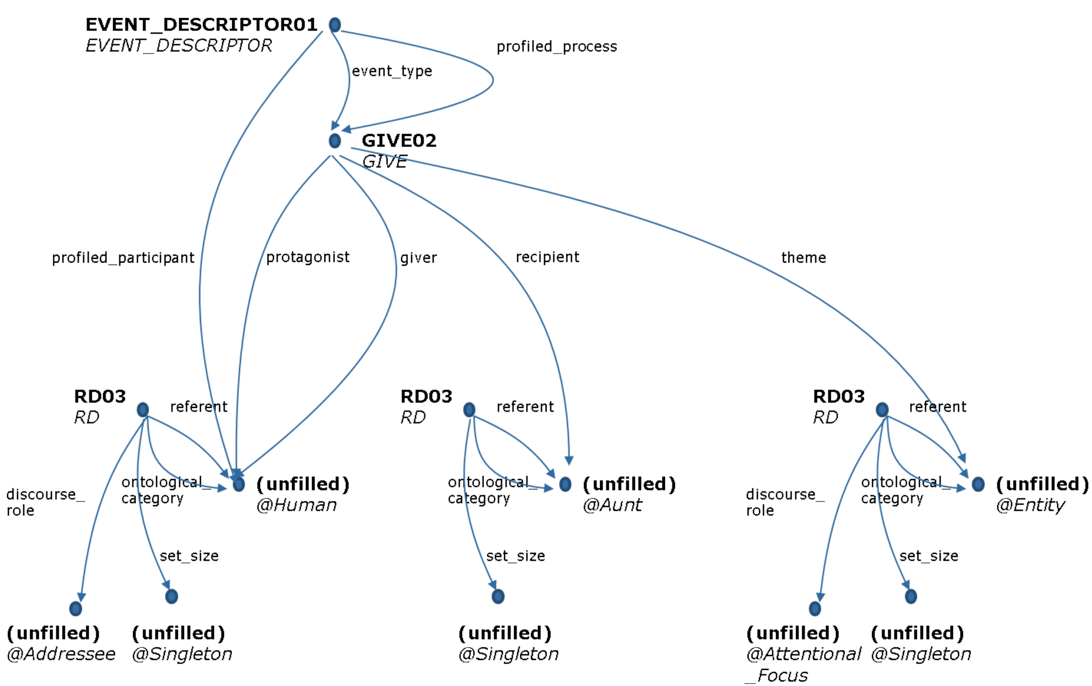

Figure 2.2 Directed Acyclic Graph representation of the semspec of *ni3 gei3 yi2* (you give aunt). Edges denote features and nodes denote slots, which can have type constraints (in italics) and atomic fillers (not shown) or structured fillers (outgoing edges from the slot).

Entities and referring expressions

Words like *ni2* (you) and *yi2* (aunt) are represented as lexical constructions and perfectly illustrates how grammar is tied to context in ECG. Both constructions are **subcase**s **of** the referring expression construction REF_EXP, which is an **abstract** construction[12] that defines a grammatical category. A referring expression denotes an @Element in context. This is written as a type constraint on its **meaning** pole, which is accessed using the special slotchain **self**.m. The **self** keyword in ECG refers to the current schema and m is short for meaning. As subcases of REF_EXP, NI2 and YI2 inherit its roles and constraints but can posit further type restrictions such as @Human and @Aunt. These types are defined in an external ontology (signified by the @ symbol in ECG and written in title case in this thesis as a convention).

```
construction NI2
   subcase of REF_EXP
   form
      self.f.orth <-- “ni2”
   meaning : @Human
   constraints
         rd.ontological_category <--> self.m
         rd.discourse_role <-- @Addressee
         rd.set_size <-- @Singleton
```

```
abstract construction REF_EXP
   meaning: @Element
      evokes RD as rd
      constraints
         rd.referent <--> self.m
```

```
construction YI2
   subcase of REF_EXP
   form
      self.f.orth <-- “yi2”
   meaning : @Aunt
      constraints
         rd.ontological_category <--> self.m
```

```
schema RD
   roles
      classifier_category
      grammatical_gender
      givenness
      definiteness
      ontological_category
      discourse_role
      set_size
      quantifier
      participanting_event
      modifiers
      referent
```

Figure 2.3 RDs, or referent descriptors, tie meanings of referring expressions to context in ECG.

[12] A newer **general** keyword replaces the **abstract** keyword in newer ECG conventions, but to avoid confusion with concrete constructions that are learned generalizations of lexically-specific constructions, the abstract keyword will be used throughout this dissertation.

The REF_EXP also ties the construction to context by providing a package of information in the RD schema (short for referent descriptor) evoked in the meaning pole. The **evokes** … **as** keyword denotes this relationship of concept co-activation by specifying the type and local name of an evoked item. The RD schema contains relevant syntactic and semantic features for resolving references to context; its presence in a constructional analysis signals that information about an entity needs to be retrieved from context. In the current model (Bryant, 2008a), the analyzer looks in its recency model for entities that match those specified features, the exact collection of which is customizable for the application. Examples of syntactic features are classifier_category, grammatical_gender, givenness, and definiteness, and examples of semantic features are ontological_category, discourse_role, set_size, quantifier, participanting_event, and modifiers.

The current learning model use a much smaller subset of these features, namely, ontological_category (a type defined in the external ontology, usually specified by something like common nouns), discourse_role (the role in the current segment of discourse, denoted by one of: @Speaker, @Addressee, and@Attentional_Focus), and a set_size (the number of the referent, denoted by one of : @Singleton, @Pair, and @Multitude). In the NI2 construction, the combination of the constraints highlighted in bold in Figure 2.3 suffices to direct the analyzer to look for a singular addressee in the current discourse segment. It also specifies that the addressee has to be (construable as) a @Human. In contrast, the YI2 construction does not require a particular discourse role of its referent, so the analyzer is directed to look for some entity in context of type @Aunt.

Actions and verbs

Figure 2.4 has an example GIVE schema that denotes the prototypical giving scene and a lexical construction GEI3 for the polysemous Mandarin morpheme *gei3*. Each sense of *gei3* is written down as a separate construction; the one shown here has the meaning of GIVE which involves three participants (reflecting a child's understanding of physical giving).

Notice that GIVE is a **subcase of** the TRANSFER schema (which is in turn a subcase of ACTION which is a subcase of PROCESS, not shown). Schemas, like the constructions and the ontology, are structured by an inheritance lattice, supporting all the expected inheritance semantics along with multiple inheritance. The semantic hierarchy plays a crucial role in generalization by guiding the formation of grammatical categories, as will be shown in later chapters. Both typed and untyped **roles** can be specified for each schema. In this case, the TRANSFER schema inherits a protagonist role from the ACTION schemas, specifies three additional roles (giver, recipient, and theme), and furthermore unifies the protagonist with the giver role through an identification constraint[13] (shown in bold). The GIVE schema in turn constrains these roles to be Animate, Animate, and Manipulable_Inanimate_Object respectively.

Another important idea to this research is that of core semantic roles, which may or may not be required as syntactic arguments. Core semantic roles, long recognized in PropBank (Kingsbury & Palmer, 2002) and FrameNet (Baker, Fillmore & Lowe, 1998), can be roughly characterized as "central participants in the conceptualization of an event". In a TRANSFER scene, the giver, recipient, and theme roles are all core roles – meaning that there must be three parties involved, even if some of them may remain unnamed. A notion of core roles is important for understanding non pro-drop languages in cases when some constructions make it optional for

[13] also called a unification constraint or a coindexation constraint.

expressing some core roles (e.g. the seller is implicit in *I bought the car for $1500.*), but it is even more crucial in pro-drop languages when core roles are often unexpressed. The learning model must be aware of the core roles in order to know which semantic pieces to recover from context, and thus to realize which constructional constituents are omissible.

Core semantic roles are not explicitly marked in ECG schemas and are given by schema-writing convention (local roles are generally taken as core). In an adult grammar, different verbs put different demands on how core arguments are expressed(e.g. *buy* requires the buyer and goods to be expressed, while *sell* requires the seller and goods to be expressed). By convention one RD per core role is evoked by verbs to notate such differences[14]. Obviously, this is part of the grammatical knowledge a learner has to acquire and the model is not supplied with these verb-specific RDs in its starting lexicon of verbs.

```
abstract construction VERB
   meaning: PROCESS
```

```
construction GEI3
   subcase of VERB
   form
      self.f.orth <-- "gei3"
   meaning : GIVE
```

```
schema TRANSFER
   subcase of ACTION
   roles
      giver : @Entity
      recipient : @Entity
      theme : @Entity
   constraints
      giver <--> protagonist
```

```
schema GIVE
   subcase of TRANSFER
   roles
      giver : @Animate
      recipient : @Animate
      theme : @Manipulable_Inanimate_Object
```

Figure 2.4 Processes, like other schemas, are described in a schema hierarchy and have core semantic roles (not explicitly marked in ECG). Shown here is the verb *gei3* (give), which has a meaning of GIVE.

[14] FrameNet (Baker et al., 1998) makes a further distinction between the unexpressed or Null Instantiated arguments: in the case of *I bought a car for $1500*, there is a definite, specific seller even if it is unnamed, whereas in *Have you eaten?*, a generic instance of meal is pictured. These two cases are referred to as Definite Null Instantiation (DNI) and Indefinite Null Instantiation (INI) respectively.

Events and clauses

An event is made up of one or more actions in different temporal configurations. An event descriptor characterizes events by the overall scene as well as the profiled process, profiled participant and modifiers. The separate designation of event type from profiled process provides added compositionality: an INTENDED_TRANSFER scene may have a BAKE event as its profiled process, as in *Mary baked me a cake*, or a CAUSED_MOTION may have a HIT event (which does not always entail motion) as its profiled process, as in *hitting the ball out of the ball park*.

```
schema EVENT_DESCRIPTOR
   roles
      event_type : Process
      profiled_process : Process
      profiled_participant : @Element
      event_structure : Event_Structure
      modifiers : Modification
      spatial_setting
      temporal_setting
      affectedness : @Affectedness
      referent : @Process
```

Figure 2.5 The EVENT_DESCRIPTOR schema, which characterizes finite clauses, has distinct roles event_type and profiled_process that allow a full range of complex events to be described.

Putting it together, prior to the resolution of references to context, the utterance *ni3 gei3 yi2* specifies "the addressee gives to aunt whatever it is that the speaker and addressee are jointly attending to". This is accomplished by combining a subject with a DITRANSITIVE_VP construction that specifies the syntactic and semantic relations between the words, shown in Figure 2.6.[15] Though the Mandarin ditransitive has stricter semantic restrictions on its verb, its central case is comparable to the English one. There are three blocks in this ditransitive construction: **constructional**, **form**, and **meaning**. Ignoring the numbers in brackets for the

[15] This DITRANSTIVE construction and others in this chapter are vastly simplified for the purpose of illustrating features in ECG. The full grammar written for the parsing exercise described in Section 2.5 uses a more complicated grammar hierarchy and slightly different meaning representations to enhance constructional compositionality.

moment, the constructional block specifies that there are three constituents in this construction: the double objects, which are REF_EXP with local names n1 and n2, and one VERB with a local name v.

```
construction DITRANSITIVE_VP
   subcase of TRANSITIVE_VP
   constructional
      constituents
         v : VERB
         n1 : REF_EXP           [0.6, 1.0]
         n2 : REF_EXP           [0.4, 0.7]
   form
      constraints
         v.f meets n1.f
         n1.f meets n2.f
   meaning : EVENT_DESCRIPTOR
      evokes TRANSFER as transfer
      evokes RD as rd1
      evokes RD as rd2
      evokes RD as rd3
      constraints
         self.m.event_type <--> transfer
         self.m.profiled_process <--> v.m
         transfer <--> v.m
         self.m.profiled_participant <--> transfer.giver
         transfer.recipient <--> n1.m
         transfer.theme <--> n2.m
         rd1.referent <--> v.m.giver
         rd2.referent <--> v.m.recipient
         rd3.referent <--> v.m.theme
         rd3.discourse_role <-- @Attentional_Focus
```

```
abstract construction TRANSITIVE_VP
   subcase of VERB_CLAUSE
   meaning: EVENT_DESCRIPTOR
```

Figure 2.6 The EVENT_DESCRIPTOR schema in use in the DITRANSITIVE_VP construction. Notice that the event_type is a TRANSFER scene while the profiled_process can be made more specific by the verb. In this case the transfer scene is unified with the verb meaning, but it need not be in other constructions, such as the caused motion construction.

The form block constrains the ordering of these three constituents, ensuring that the verb v appears immediately before the first object n1 and n1 immediately before the second object n2. The meaning of this verb phrase is an EVENT_DESCRIPTOR schema whose scene type is a

TRANSFER scene. This is accomplished through the first two constraint lines highlighted in bold in Figure 2.6. The profiled_process is specified by the meaning of the verb, which is constrained to be a verb that has a transfer meaning in the next bolded constraints. The core roles of the transfer scene are handled in the next three unification constraints. The profiled_participant in this event is to be supplied by the subject, but whatever it is, it is also the giver in the scene. As expected, the recipient and theme in the transfer scene are supplied by the two referring expressions. Finally, three RDs are evoked by this construction for the three core roles along with any particular contextual restrictions. For the purpose of illustration, neither the giver nor recipient is restricted, but the theme has to be the attentional focus of the current discourse.

The giver is supplied by the subject in the ACTIVE_SUBJECT_VP construction shown in Figure 2.7. Recall that the profiled_participant of the event described by the DITRANSITIVE_VP construction is already unified with the giver role, so all that the ACTIVE_SUBJECT_VP construction needs to do is to unify the subject with the profiled_participant in the VERB_CLAUSE construction.

```
construction ACTIVE_TRANSITIVE_VP
  subcase of CLAUSE
  constructional
    constituents
      s : REF_EXP          [0.3, 1.0]
      vp : VERB_CLAUSE
  form
    constraints
      s.f before vp.f
  meaning : EVENT_DESCRIPTOR
    constraints
      self.m <--> vp.m
      self.m.profiled_participant <--> s.m
```

Figure 2.7 An ACTIVE_TRANSITIVE_VP construction that handles voicing exploits compositionality in ECG.

As explained briefly in Chapter 1, the learner begins with three the abstract categories: MORPHEME, PHRASE, and CLAUSE, which roughly correspond to “a word”, “parts of an utterance", and "complete utterance". The learning model uses the EVENT_DESCRIPTOR schema as the default meaning of a CLAUSE for extensibility, but the remainder of this dissertation will skip the display of the EVENT_DESCRIPTOR schema and use the profiled_process (which in most basic cases is the same as the event_type) directly as the meaning pole of clauses for brevity.

Optional, omissible and extraposed arguments

It was introduced in Chapter 1 that two interesting features about Mandarin Chinese is its common omission of arguments and its relatively flexible phrasal structure that allows frequent topicalization and fronting of the object. These two phenomena must be accommodated by any grammar formalism. ECG includes two representational mechanisms, supported by underlying processing machinery, that handle fronting and omission. We first introduce the representation here, and discuss the processing support in Section 2.3.

As shown in the DITRANSITIVE_VP construction in Figure 2.6, two of its constituents (n1 and n2) have bracketed probabilities to the right of its type constraint. Each set of bracketed numbers for constituent β denote

[P(β is expressed), P(β is expressed locally | β is expressed)]

For short we will henceforth refer to the two probabilities as the locality probabilities. For example, the [0.6, 1.0] next to the constituent n1 indicates that the first object in the DITRANSITIVE construction is expected to appear 60% of the time and always in its specified location inside the construction and never extraposed. This is in contrast with the second object n2, which is assigned [0.4, 0.7], indicating that it is not only expected to be present 40% of the time, but is also expected to show up outside the normal DITRANSITIVE_VP construction in 30% of the cases when it is overtly mentioned. These probabilities are chosen to reflect the ability to extrapose or front the object in Mandarin, as in *ba3 tao3 gei3 yi2* ('CV$_{obj}$ peach give aunt')

The fronting construction that puts the fronted object together with the VP makes use of another ECG keyword, **extraposed**. While constructions with an extraposed constituent are not learned by the current learning model as implemented, they can be with an easy extension to the model. More importantly, these constructions are introduced here to demonstrate the representational power of ECG and its suitability as a grammar representation for learning models. In the TRANSITIVE_VP_WITH_FRONTED_OBJECT construction shown in Figure 2.8, the object is marked with the **extraposed** keyword. During processing, the extraposed object sets up a syntactic context in which later constituents (in this case the TRANSITIVE_VP construction) can have one of its constituents expressed non-locally.

```
construction TRANSITIVE_VP_WITH_FRONTED_OBJECT
   subcase of VERB_CLAUSE
   constructional
      constituents
         ba : BA3-CV
         extraposed obj : REF_EXP
         vc : TRANSITIVE_VP
   form
      constraints
         ba.f meets obj.f
         obj.f meets vc.f
   meaning : EVENT_DESCRIPTOR
      constraints
         self.m <--> vc.m
         ba.prototransitive <--> vc.prototransitive
         ba.prototransitive.proto_patient <--> obj.m
         vc.m.affectedness <-- @Affected
```

Figure 2.8 An object-fronting construction in Mandarin demonstrates the use of the extraposed keyword. It sets up an expectation that one of the constituents of the vc (TRANSITIVE_VP) is expressed nonlocally.

Unlike omissible or extraposed constituents, which generally constituent the core parts of the meaning, some constituents are truly optional. These are sometimes referred to as adjuncts but in this work are inclusive of modifiers (nominal and clausal) and sometimes grammatical morphemes. These are marked in ECG with the **optional** keyword and can have an associated probability of being expressed. For example, an optional post-verbal aspect marker may be added as a constituent of the DITRANSITIVE_VP2 construction as shown in Figure 2.9, with a locality probability P(β is expressed) specified at 0.5.

The probabilities shown in the examples are based on a grammar writer's intuitions. As a proof of concept, I carried out an exercise jointly with John Bryant in learning these parameters about a CHILDES corpus in Mandarin Chinese. Those results are described in Section 2.5.

```
construction DITRANSITIVE_VP2
   subcase of TRANSITIVE_VP
   constructional
      constituents
         v : VERB
         optional post_asp : POSTVERBAL_ASPECT_MARKER      [0.5]
         n1 : REF_EXP          [0.6, 1.0]
         n2 : REF_EXP          [0.4, 0.7]
   form
      constraints
         v.f meets n1.f
         n1.f meets n2.f
   meaning : EVENT_DESCRIPTOR
      evokes TRANSFER as transfer
      evokes RD as rd1
      evokes RD as rd2
      evokes RD as rd3
      constraints
         self.m.event_type <--> transfer
         self.m.profiled_process <--> v.m
         transfer <--> v.m
         self.m.profiled_participant <--> transfer.giver
         transfer.recipient <--> n1.m
         transfer.theme <--> n2.m
         rd1.referent <--> v.m.giver
         rd2.referent <--> v.m.recipient
         rd3.referent <--> v.m.theme
         rd3.discourse_role <-- @Attentional_Focus
```

Figure 2.9 A DITRANSITIVE_VP2 construction that allows probabilistic omission of its objects as well as an optional postverbial aspect marker.

Discourse and speech acts

One final aspect of the grammar representation in this model is the description of a segment of discourse. The DISCOURSE_SEGMENT schema, often shortened as DS, provides a simple way of notating the identity of the speaker, addressee, attentional focus, and speech act. One DISCOURSE_SEGMENT schema is attached to the ROOT of each analysis to help ground the analysis in context.

Speech acts are also represented as schemas. The general SPEECH_ACT schema has roles for the speaker, the addressee, the content of the speech act as well as the forcefulness of the tone. Speech acts are further divided into their subtypes: *explaining*, *answering*, *approving*, *admonishing*, *requesting action*, *requesting answer*, *calling*, *exclaiming*, and *practicing*. Each is represented as a schema with roles that further elaborate the interchange. More about why these schemas are chosen and how they are used to annotate the learning data will be described in Chapter 7.

```
schema DISCOURSE_SEGMENT
  roles
    speaker : @Human
    addressee : @Human
    attentional_focus : @Entity
    speech_act : SPEECH_ACT
  constraints
    speaker <--> speech_act.speaker
    addressee <--> speech_act.addressee
```

Figure 2.10 The DISCOURSE_SEGMENT schema is attached to the root of each analysis.

```
schema SPEECH_ACT
  roles
    speaker : @Human
    addressee : @Human
    content : Event_Descriptor
    forcefulness : @Forcefulness
```

Figure 2.11 The SPEECH_ACT schema gives additional details about the speech act, including the content and forcefulness (as indicated by intonation). Subtypes of speech acts elaborate on the details of the speech act.

2.2 Simulating events in context to update the world model

A rich context model is important for understanding language, particularly when arguments are omitted from utterances and have to be retrieved from context. The last section discussed the mechanisms (RD, or referent descriptors) with which language is tied to context. This section describes how the context model is built and maintained dynamically in order to assist the understanding of child and parent utterances and bootstrap grammar learning.

A child learner has access not only to visual, audio, and tactile input during her daily interaction with her caretaker, but also to her internal states and desires (e.g. being hungry, wanting to eat, or wanting to be held). These sorts of situational information undoubtedly factor into language learning in unexpected ways but are beyond the scope of the current context model. There are, however, two basic functions that the model of context must provide:

1. Remembering a recent history of mentioned and enacted actions so that linguistic references to entities and events can be resolved against context.
2. Tracking the state of people and objects in the surrounding (e.g. locations, postures, and possessions), so that the plausibility of future events can be evaluated.

The current system relies on a simplistic relational database to store facts about entities and relations per time slice and provide limited capability of maintaining knowledge through time. The details of how the context model is implemented are irrelevant for current purposes except for the basic structure of the context model shown in Figure 2.12. The model is organized by temporal durations into intervals. Contained in intervals are context elements and relations on these context elements. Roughly speaking, noun phrases refer to context elements whereas finite clauses and verbs refer to intervals. Intervals can be contained inside another interval, as shown by downward arrows in the diagram: all relations that are true in the parent interval are expected to hold true in the sub-interval. Horizontal arrows are temporal links representing propagation through time: only a subset of (predefined) relations persists through time.

The top-most interval, `Dialogue`, represents the duration of the entire dialogue and records facts that hold true throughout the whole dialogue, such as who the participants are, what objects are around, and relations between them. Other relations such as locations of objects change throughout time and are recorded in the `Start` interval. As expected of a transcript of parent-child interaction, there are interleaving events and utterances (labeled `DS` for discourse segment). Events can have sub-components and these are represented using the same sub-interval mechanisms. For any given utterance, the previous history of events and discourse segments leading up to it are available to the analyzer and the learning model. The next two sections describe how the system utilizes this information to aid language understanding. The rest of this section gives a quick sketch on how the context model is dynamically updated throughout the course of a dialogue from the transcribed learning data.

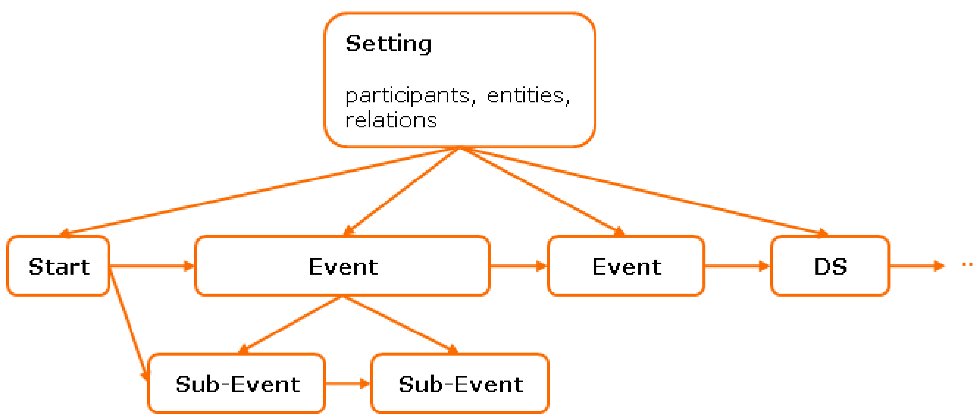

Figure 2.12 The context model supports the analysis process by tracking sequences of events and dialogues and is arranged hierarchically using temporal intervals. Relations in a parent interval (e.g. Setting) hold true in a child interval (e.g. Start) unless explicitly blocked. This model of context is used to maintain (non-probabilistic) beliefs over time and is updated using event annotations in the learning corpus and a simple simulator.

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

MOT: ni3 rang4 a1+yi2 chi1 (you let aunt eat)

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

MOT: ni3 gei3 yi2 (you give aunt)

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

INV: xie4 xie4 (thank you)
```

Figure 2.13 Utterances and event annotations surrounding the *ni3 gei3 yi2* utterance. [16]

The context model contains information about ongoing events, and at the start of a dialogue, the context model contains only the `Setting` and `Start` intervals. The contents of these intervals and the subsequent event and discourse intervals have to be created from the input data, which are transcribed dialogues. Unfortunately such rich annotations are not available as part of the Beijing CHILDES corpus and had to be added manually. The added annotations include information about the surroundings (e.g. objects within the room), initial settings, and ongoing events. With no video or audio transcript available, event annotations are inserted whenever an

[16] Compound words such as *a1+yi2* come pre-segmented as part of the CHILDES corpus and most are left intact in the learner's lexicon as fixed chunks . Other pre-segmented nouns such as *xi1+gua1* (watermelon) and verbs such as *hao3+kan4*(pretty, lit. good looking) or *liu2+xue3* (bleed, lit. stream blood) are left as compound words in the learning input as well as the starter lexicon. They seem to be collocated frequently enough that they should be made available to the learner as fixed phrases. Exceptions are verb + resultative particle compounds such as *zou3+kai1* (walk away) and *jian3+qi3+lai2* (pick up, lit. pick up towards) which are deemed inappropriate as pre-existing knowledge for the learner. The pre-segmentation is thus removed and the learner has to learn the combinations.

action is judged to have taken based on the ongoing dialogue. Any reasonable action in each situation suffices for the purpose of modeling grammar development. It is important to note, however, that events are annotated according to their times of occurrence. This means that the events mentioned in utterances, especially commands, are not always found in context.

Example event annotations, as inserted around the utterance *ni3 gei3 yi2* (you give aunt), are shown in Figure 2.13 in XML (see Appendix B for the complete transcript). Each event is specified by an event type and a unique ID[17] . Highlighted in bold in Figure 2.13 is the event `fetch01` of category `Fetch`. The annotated event types are drawn from the ontology, whose hierarchy is assumed to mirror that of the schema process hierarchy (recall that the ontology is expected to contain linguistically relevant information and more). Participants are specified on the event, such as the fetched role being filled by the peach[18].

Because the learning model requires that the context model not only record events but also track the entities, each piece of event annotation is further processed by a simulator that updates the context model with consequences of the action. Compatible with the idea of simulation-based understanding but drastically simplified for current purposes, the simulator uses a given set of scripts to make inferences about pre-conditions and post-conditions. For example, at the end of the `Offer` event, the child retains possession of the peach, whereas at the end of the `Give` event, the investigator obtains possession of the peach. Though this is not currently done, the same simulator can be used to evaluate different interpretations of an utterances by checking if mentioned events have satisfiable pre-conditions and can be carried out.

2.3 Finding the best-fit analysis of an utterance given limited context

[17] specified in XML as `event` objects with the properties `cat` and `id` respectively

[18] specified in XML as `binding` objects using properties `field` and `ref`, respectively.

The ability of learning model to learn from a situated utterance depends greatly on its ability to determine the most appropriate interpretation of the utterance. Effortless as language comprehension may seem in adults, this is no simple task due to inherent ambiguity in natural language. There may be multiple interpretations of an utterance even when constructions have detailed form and meaning constraints. Grammar, semantic judgment, situational and discourse context as well as conventional usage all play a role in this selection process.

Consider the running example, *ni3 gei3 yi2* (you give aunt). The strict ordering constraints and type restrictions (e.g. that the theme must be a Manipulable_Inanimate_Object) of the DITRANSITIVE_VP leaves little room for ambiguous interpretations. However, more generous type restrictions (e.g. that the theme can be any Physical_Object, including Human) are expected in a more sophisticated, adult-like grammar. In this case, there are at least two interpretations of the sentence *ni3 gei3 yi2*, one where the aunt is analyzed as the recipient, and the other where the aunt is the theme. It may be semantically unreasonable for an aunt to be a theme and much more reasonable for her to be a recipient, and this is exactly the sort of information that needs to go into choosing the correct analysis. Another less obvious source of information has to do with grammar usage statistics — how often the theme is omitted versus how often the recipient is omitted, as well as how often the first object of the ditransitive is filled by a pronoun versus a common noun or a proper noun, etc. A third source of information comes from the situational context surrounding the utterance: if the mother is gesturing at a peach, or if the child has just given the peach to the aunt, then there is an obvious better interpretation of the utterance.

A language understanding system needs to concretize these intuitions and the learning model uses Bryant's best-fit constructional analyzer (Bryant, 2008a) to determine the best interpretation or analysis of each utterance in context. Best-fit refers to a quantitative measure of integrating multiple sources of information, in this case constructional, semantic, and contextual cues. Context-free grammar formalisms have long been pragmatically extended with probabilities on rules, referred to as Probabilistic Context-Free Grammars (PCFG) in traditional parsing systems. The idea is similar in construction grammar, but probabilities can be attached to both constructional constituent as well as semantics roles. Broadly speaking, the task of the analyzer is to find the most probable analysis given a grammar, a sentence, and its context:

$$a = \underset{a}{\operatorname{argmax}} \mathrm{P}(a \mid sentence, grammar, context)$$

The Bryant analyzer uses a left-corner parsing algorithm which maintains a stack of possible, competing analyses as it incrementally processes the input. The reader can refer to (Bryant, 2008a) for how the analyzer decomposes the above probability into manageable factors that can be estimated incrementally, but the following example appeals to the intuition behind the parsing algorithm. As the analyzer encounters each word in the input, it tries its best to connect it to the current best-guess constructional tree, as illustrated in Figure 2.14. After the first two words in *ni3 gei3 yi2*, one of the several competing analyses in Figure 2.14 looks like a fragment of the ideal analysis in Figure 2.1: the analyzer has built up a partial structure over the *ni3* and *gei3*, and now needs to decide what to do with *yi2*. It needs to **push** a lexical construction onto the stack to cover the word, and in this case it is straightforwardly YI2 (this may not be so easy a decision if there are homonyms in the grammar). The analyzer can then **attach** YI2 to the DITRANSITIVE_VP as constituent n1 (semantically the recipient), or it can decide that n1 is omitted and attach YI2 as constituent n2 (semantically the theme).

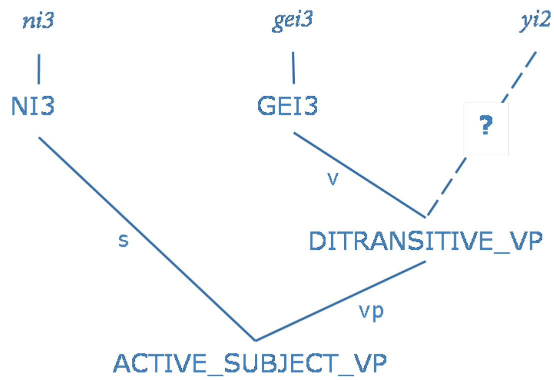

Figure 2.14 Incremental parsing: the analyzer maintains a stack of possible analyses and decides the best parsing operation to take at any given point. One possible analyses after the first two words is shown above, and the ideal analysis (with the Aunt filling the recipient role of the GIVE schema) can be achievable if the right steps are taken. Another competing analysis going forward includes the one where the Aunt fills the theme role of the GIVE schema. Other competing analyses at this stage of analysis include one where *ni3* is the subject and *gei3* is a benefactive coverb.

Constructional, semantic, and contextual cues guide the choice between various operations, and the Bryant analyzer computes the cost of each operation based on these cues. For the purpose of the current discussion, four sets of parameters in particular are relevant:

- The **locality probabilities**, i.e. $P(expressed_\beta \mid \alpha)$ and $P(local_\beta \mid expressed_\beta, \alpha)$, where both *expressed* and *local* are binary variables. This is the set of probabilities given in brackets in the ECG notation discussed in the last section: the probability that a constituent β of construction α is expressed (as opposed to omitted), and the probability that the constituent occurs locally if it is expressed.
- The **constructional filler probabilities**, $P(filler_\beta \mid expressed_\beta, \alpha)$. This refers to the probability that a constituent β of construction α is realized as $filler_\beta$ if the constituent is expressed. An example is the probability that the constituent v in the DITRANSITIVE_VP is filled by the construction GEI3 (as opposed to other verbs, for example). Given this powerful piece of statistical information, constructional

categories can technically be defined purely probabilistically as distributions over constructional fillers. However, we contend that with this approach, the number of parameters associated with the grammar is greatly increased and important linguistic insights are lost, so in this work both concrete and abstract constructions are used.

- The **semantic presence probabilities**, $\mathrm{P}(filled_{role} \mid role_f, f)$. This refers to the probability that a particular $role_f$ of a frame f is filled. This probability is particularly important for properly evaluating the semantic fit for infrequently filled roles, such as a non-core argument like instrument.
- The **semantic role probabilities**, $\mathrm{P}(role_f \mid filler, f)$. This refers to the probability that a particular filler *filler* fills a role $role_f$ in the frame *f*. One example of such a probability is that of the aunt filling the recipient role of the GIVE frame. Obviously, in a unification grammar representation of deep semantics, the same filler may participate in multiple frames. For example, the giver in a GIVE frame may also be the force_supplier in a FORCE_APPLICATION frame if a physical handing-off of an object is described. In practice an aggregating function such as averaging is used to determine the score of a particular semantic binding.

The last two probabilities make up the **semantic model** in the system, providing the commonsense or typicality judgment of events. Using probability measures derived from these parameters, the analyzer examines the trade-offs between interpretations and chooses the most likely analysis. The returned analysis consists of a constructional tree that describes constituency relations, a SemSpec containing meaning bindings, as well as a set of proposed referents for each RD. In addition, the model of context is able to provide even richer structure than the analyzer is designed to retrieve from context, and this extra post-processing step, termed context fitting, is described in the next section.

From a learning model point of view, it is important to distinguish which kinds of statistical knowledge can be presumed of the learner in the initial stage. The semantic model, assumed to be built up through interaction with the world and experience with events, is pre-existing knowledge for the learning model. On the other hand, the other two sets of parameters, the constructional filler probabilities and locality probabilities, are linguistic parameters and have to be learned. Chapter 6 is devoted to describing how the learner learns these and other statistics on the input.

Robust parsing

One crucial but yet unmentioned topic is that of analysis under noise, such as disfluencies in real speech or an incomplete or incorrect child grammar. Under such conditions the analyzer cannot expect to achieve a well-formed single-rooted constructional tree to span the entire input. Instead it tries to recognize as many coherent fragments as possible. Again an appeal to intuition is used here, and the reader is strongly encouraged to get the details from Chapter 8 in (Bryant, 2008a).

Assume that the learning model has acquired a lexically-specific chunk *ni3 gei3* by rote memorization, creating a NI3-GEI3 construction with only two constituents. When the learner now attempts to analyze *ni3 gei3 yi2*, a possible analysis after the first two words look like part (a) of Figure 2.15. The analyzer is able to use the learned construction NI3-GEI3 to cover the first two words but it now has two choices going forward: it needs to find a way to connect *yi2* to NI3-GEI3 or give up on it and make a separate constructional tree beginning with *yi2*, as illustrated in (b) and (c) respectively. In order to connect *yi2* to NI3-GEI3, it needs to **propose**[19] another construction that has both as constituents — this is shown in (b) as the bottom-most question mark. This is obviously not feasible if no such construction exists, but there are associated costs even if it does. On the other hand, the analyzer can choose to abandon (at least temporarily) the current constructional tree and start a new one at the next word, as in (c). However, the analyzer has to pay a user-set penalty for each additional root in the analysis so that the cohesiveness of the analysis is properly traded-off with the cost of using an possibly infrequent or unlikely construction to put the words together. The likely result of (c), given that *yi2* is the last word in the utterance, is an analysis with two roots (NI3-GEI3 and YI2) with no semantic connections between them.

In fact, the current analyzer has the ability to put a current constructional tree on hold, start and finish a new constructional tree, and then go back and continue with the first tree, essentially skipping over a root. This is extremely useful for analyzing disfluent speech, such as *I went to the, uh, store yesterday*, where a complete interpretation is achievable using the regular self motion construction despite an intervening chunk of interjections. It is also helpful in recovering a coherent interpretation if the grammar is limited in coverage. Consider the sentence *Put the frog on the napkin in the box*. If the grammar has no coverage for reduced relative clauses and the analyzer does not have the ability to skip roots, the only possible interpretation is one where the goal of the put action is the napkin. Using the root skipping feature, a competing analysis — the correct one, where the goal is the box — can be derived.

[19] The term **propose** , like **push** and **attach**, is used in the technical terminology of left-corner parsing. It indicates the creation of a new stack state and does not involving grammar learning as described in this dissertation.

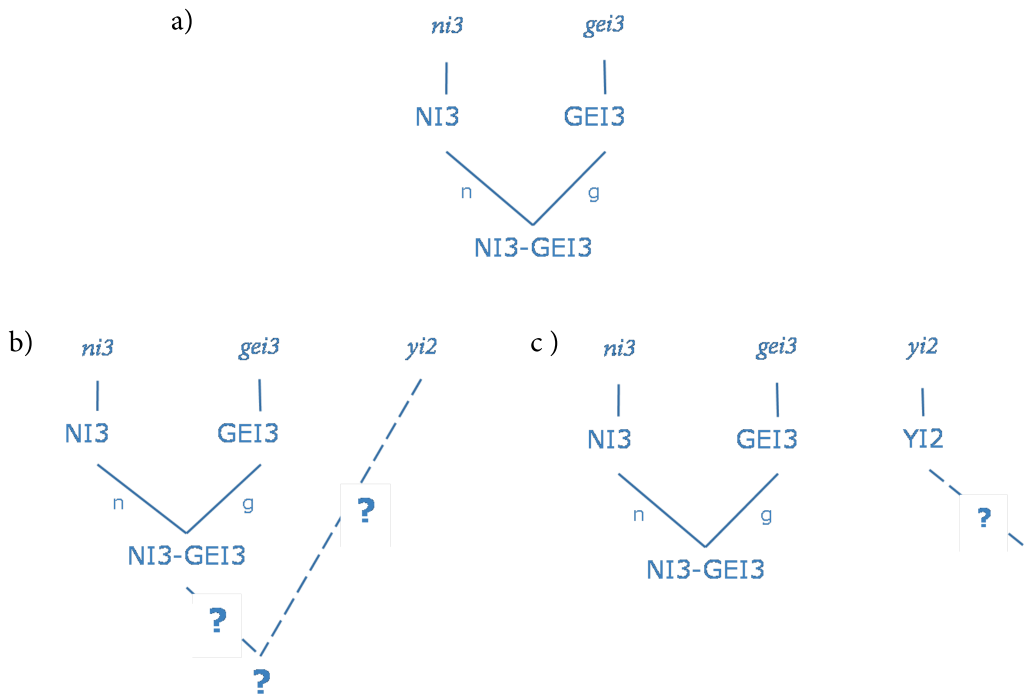

Figure 2.15 Robust parsing: in a situation with noisy input (e.g. disfluencies or incomplete grammar), the analyzer may (a) try to create a single-rooted analysis regardless or (b) decide to not connect the next word to the current constructional tree and instead start a new constructional tree for the next word, paying a penalty in the process. The resulting analysis of (b) is one with multiple constructional trees, also referred to as a constructional forest or a multi-rooted analysis.

The learner model relies critically on the analyzer's robustness feature since it enables the learner to understand utterances even without a complete grammar. From these analysis fragments the learner can create new form-meaning mappings based on relations found in context (e.g. that the aunt is the recipient in context enables a new construction to be composed between NI3-GEI3 and YI2). The recovery of meaning from context is the subject of the next section.

2.4 Fitting the best analysis to full context

Just as REFERENT_DESCRIPTORs, or RDs, are designed to capture linguistic references to entities in context (or reified processes, e.g. *an accident* or *the robbery*), EVENT_DESCRIPTORs are created to capture linguistic descriptions of events and maintain their correspondences to events in context. Unfortunately, the current implementation of the analyzer is unable to take full advantage of the context model and returns as a result only a type-compatible list of candidates for each RD, leaving events unresolved.

The main thrust of this work, on the other hand, is about how the rich structure of events in context can be used to bootstrap grammar learning, so a post-processing of the analyzer's output is necessary. After obtaining a best-fit analysis of an utterance, the learner currently performs a greedy search over the context model to match mentioned processes to prior events. In top-down order established by the precedence below, slots are searched recursively from an event instance to its roles in order to preserve the structural integrity of events. This post-processing step is referred to as context-fitting in this work.

```
Context fitting precedence:

  DS > Discourse Participant > Complex Process >
  Structured Simple Process > Unstructured Simple Process >
  Structured Entities > Unstructured Entities
```

Casual observation of the Beijing corpus and of the particular sets of dialogues chosen as learning input for this model found that parents rarely label actions as the actions unfold. Most action descriptions appear in the form of announcing an intention (e.g. *Let me get you some rice*) or a request for the child to perform an action (e.g. *Remember to spit out the bones*). Naturally, then, complete match of event descriptors to events *preceding* an utterance are infrequent. Though children may be able to use post-utterance situational context to infer word and utterance meaning, to be conservative, the learner relies only on events that occur temporally before the utterance, leading to two other forms of context matches that may be exploited:

- partial matches: given the repetitive nature of parental utterances and play sessions, the same event may often be described multiple times using different lexical items (e.g. *pick up* and *take*) or from different perspectives (e.g. *give* and *take*). Similarly, similar actions may be performed on multiple objects (e.g. picking up a bear and picking up a monkey). This kind of repetition can be exploited by allowing partial matches of an event to context as long as the types of event are closely related according to the schema hierarchy. Noise is unavoidably introduced into the learning model through partial matches, but we believe this to be consistent with mistakes a child may make in reading other people's intention.
- precondition matches: one action often enables another action to be carried out. This is the notion of preconditions in action planning in classic Artificial Intelligence. The picking up of a toy, for example, enables the child to then give it to her caregiver. The same action obviously also enables the child to then throw the toy onto the ground, so one action by itself may not be a good predictor of the next one, but it can be used to reduce the noise in the partial matching process. Though not current supported, the use of precondition matching is an easy extension to the current simulator implementation.

At the end of the context-fitting process, each slot is assigned a context element or event, or null if no match is found. In the multi-rooted analysis from Figure 2.15 (b), there are two separate slots for the recipient role of the GIVE schema and the meaning pole of YI2 ('aunt') since no semantic bindings are specified grammatically. However, after context fitting, the investigator (i.e. aunt) is found to be an appropriate contextual filler of both slots. The discovery of new semantic relations from context is an important concept for the composition operation that puts together new constructions, which will be discussed in Chapter 4.

2.5 Analyzer Demonstration: analyzing Mandarin Chinese

Now we turn to an attempt at analyzing a corpus of parent-child interaction in Mandarin Chinese using a hand-written grammar in hopes of demonstrating both that the chosen grammar formalism is sufficiently powerful to capture linguistic knowledge pursued by the learning model and that the analysis mechanism is able to generate the desired analyses in the ideal grammar scenario. There are two parts to this exercise, the first of which is an evaluation of the constructional analyzer on the child language corpus using the handwritten grammar, and the second of which is an attempt to learn some parameters associated with the grammar in a supervised way. This work, carried out in collaboration with John Bryant and updated for this dissertation, has also been described in (Bryant, 2008a) with emphasis on algorithmic extensions to the analyzer to provide robust analysis in conditions of incomplete grammar and disfluencies in natural speech.

The data consists of 35 short dialogues and 4 long dialogues taken from the Tardif Beijing corpus in CHILDES (MacWhinney, 2000; Tardif, 1993; Tardif et al., 1997). Data from four subjects make up these dialogues. Short dialogues are manually selected on the basis of forming a coherent episode of interaction. Selected utterances are spaced closely together temporally in the transcript and contain generally no more than two ongoing activities, e.g. eating dinner and child being distracted with some toys, whereas the long dialogues are contiguous segments of transcripts taken from the beginnings and ends of recording sessions. Each short dialogue contains on average 11 clauses and each long dialogue contains on average 80 clauses. Of those

705 clauses, 132 contain only interjections and are not used for evaluation purposes. The mean length of the remaining 570 content clauses is about 3.58 words. Clauses are obtained by separating transcribed utterances wherever pauses are notated in the transcript; this is done for convenience and uniformity since naturalistic data often contains run-on sentences of different speech acts. For the purpose of the parameter learning exercise, the 35 dialogues and 2 of the long dialogues are used as training data and the remaining 2 long dialogues are used as test data.

In order to automatically evaluate the correctness of the analyses, frame-based semantic "gold standard" annotations are manually added for each utterance. The semantic annotations are added for both verbal arguments as well as argument structure arguments, as shown in Figure 2.16 for the sentence *mo3 wai4+tou2 ke3 jiu4 bu4 hao3+kan4 le a* ('if you apply [the lotion] to your forehead then you won't be pretty').

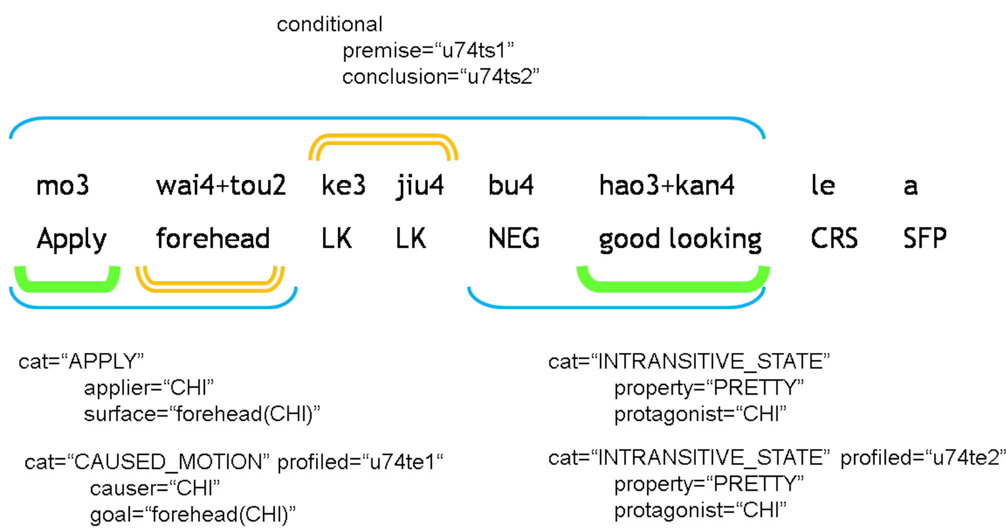

Figure 2.16 Gold standard annotation of the utterance *mo3 wai4+tou2 ke3 jiu4 bu4 hao3+kan4 le a* ('if you apply [the lotion] to your forehead then you won't be pretty'). Both verb arguments and argument structure (phrasal) arguments are annotated, as shown in the bottom four annotations for the two clauses. Bracketing information is supplied (verb brackets shown in thick solid lines, phrasal brackets in thin solid lines, and any additional words or arguments in double-lines) as well as any interesting sentential constructions (e.g. conditionals).

As shown in the bottom half of the diagram, each verb phrase is annotated with two layers of annotation: that the verb *mo3* has a meaning of APPLY, whose arguments are the child and the child's forehead, and that the phrase *mo3 wai4+tou2* has a meaning of a CAUSED_MOTION scene whose arguments, as it turns out, overlap completely with the verb arguments. The same is done for the second half of the utterance. These four annotations in the bottom are the bulk of the scoring criteria used for the output of the constructional analyzer. The correctness of the event type as well as the argument types are counted in the core argument scores, with minor adjustments for argument omission. Further details of the scoring algorithm are given in Chapter 7.

For annotation completeness, additional linguistic phenomena are noted, e.g. the conditional with the first event (applying lotion to the forehead) as the premise and the second event (not being pretty). This information, however, is not used in the final scoring due to the inadequacy of the handwritten grammar[20].

The Mandarin Chinese grammar is written based on native speaker's intuition and the short dialogues alone. The idea is that a reasonable grammar should extend with statistical information to cover unseen data. The grammar has 263 schemas for common event types (e.g. eating, playing, throwing, being naughty) and an additional 174 ontology types for common entities (e.g. doll, soup, rice, chairs). These are paired with about 609 constructions of which:

- a few are blanket constructions for handling unknown words, interjections and reduplication
- 106 are abstract and define phrasal and lexical category structure.

[20] As it turns out, conditionals and other clausal relations in Mandarin are very difficult to deduce grammatically. Often conditionals are implied by temporal conjunctions, and temporal relations are implied by phrasal ordering. Those interested can refer to (Yang, 2007) for an account of conditional constructions in Mandarin.

- 73 are concrete phrasal or lexical constructions including argument structure constructions, noun phrase constructions, modifier constructions and a few sentential constructions.
- 283 are open class lexical items such as verbs and nouns (covering all dialogues)
- 135 are closed class function words

Analyzing the corpus with the handwritten grammar

Without learning any additional statistical parameter except for the grammar writer's expected omission probabilities of arguments, the analyzer is used to analyze the 700 or so utterances *without* using the robustness feature. About 136 utterances fail to be analyzed due mostly to insufficient coverage of the grammar. Of the remaining 560 or so utterances in the training and test sets, the analyzer achieves the results reported in Figure 2.17 using the automatic scoring.

	Core Argument Precision	Core Argument Recall	Core Argument F-score	Resolution F-score
training set (35 short + 2 long)	0.824	0.744	0.782	0.685
test set (2 long dialogues)	0.819	0.724	0.768	0.785

Figure 2.17 Automatic scoring of the analyses of child-directed utterances. Core argument scores are based on the type constraints given in the SemSpec, whereas resolution score are based on the resolution results (entities in context) suggested by the analyzer.

The analyzer is expected produce a different number of semantic bindings than is given in the gold standard annotation, so a modified precision/recall measure is used to score the returned analyses, the details of which are given in Section 7.2. Core arguments are scored on the basis of the type constraints given in the SemSpec, whereas resolution results are scored based on the context elements suggested by the analyzer for each RD.

As can be observed, the analyzer does a reasonably good job with the utterances. It is difficult, however, to assess these scores without a standard benchmark. I have therefore performed a detailed manual analysis of the first 150 utterances in the training set ignoring repeats and this result is reported in (Bryant, 2008a) as well. The analyzer found an analysis for 125 of the 150 utterances; the remaining 25 required constructions that were not part of the handwritten grammar. Of the 125 returned analysis, 76 of which were judged to be the correct analysis in terms of both the SemSpec and the resolution results. Of the 49 incorrect analyses,

- 8 had the right constructional interpretation but the omitted arguments were incorrectly resolved,
- 9 could not be properly analyzed because the necessary constructions were not in the grammar. The analyzer instead creatively used a combination of other constructions to interpret the utterance,
- 12 used an incorrect word sense or had an incorrectly attached modifier phrase,
- 3 had problems with topicalization,
- 12 had incorrect constituent omissions,
- 1 had a problem with reduplication, and
- 3 had trouble with a sentence final *le* marker, which sometimes act as an aspect marker and sometimes a current relevant state marker and sometimes both.

Tuning the statistical parameters of the grammar

Finally, we describe a brief exercise in trying to learn from data more accurate locality probabilities, constructional filler probabilities, and semantic role probabilities. Using the supervised, iterative estimation techniques described in (Bryant, 2008a), the returned analyses of the training set are automatically re-ranked using the gold standard and the probabilities are recalculated using the post-reranking top analyses. The learned parameters are then used in analyzing the test set.

This exercise is a proof-of-concept in testing the learnability of the parameters. The training data is very limited in view of the number of parameters that need to be estimated, and is certain very small compared to most other machine learning applications in NLP systems. Nonetheless, an encouraging improvement of parsing performance is observed, as shown in Figure 2.18. These results are updated from the ones described in (Bryant, 2008a) due to slight grammar and scoring changes but show the same trends. We take these results to be an indication that this may be a fruitful direction for future research.

Training set	Core Argument Precision	Core Argument Recall	Core Argument F-score	Resolution F-score
0	0.824	0.744	0.782	0.685
1	0.843	0.798	0.820	0.747
2	0.843	0.798	0.820	0.745
3	0.844	0.798	0.820	0.741
4	0.844	0.798	0.820	0.742
5	0.845	0.798	0.821	0.745

Test set	Core Argument Precision	Core Argument Recall	Core Argument F-score	Resolution F-score
0	0.819	0.724	0.768	0.785
1	0.882	0.824	0.852	0.825
2	0.881	0.815	0.846	0.838
3	0.881	0.815	0.846	0.831
4	0.881	0.815	0.846	0.838
5	0.881	0.815	0.846	0.831

Figure 2.18 Iterative estimation of grammar parameters lead to improvements in both the training and test set. Iteration 0 is the initial state without parameters as replicated from Figure 2.17. The learned parameters surprisingly lead to greater improvements in the test set, and could be due to the fact that the grammar was written with the training set closely in mind.

Chapter 3.

Learning a Construction Grammar

This chapter gives the first technical overview of how the learner creates new grammatical structures that link form relations to meaning relations. A stable starting vocabulary is assumed for the learner since it is outside the scope of the current thesis to address word learning directly, but it will become clear as we go along that word learning is extremely compatible with the current framework. We will revisit the topic of word learning in the final chapter; for now we will turn our attention to how grammatical structures are learned by the model.

Readers who are familiar with chunk-and-merge style grammar induction in context free grammars e.g. (Langley & Stromsten, 2000; Wolff, 1988) can relate easily to the basic operations proposed here. **Composition** is the basic chunking mechanism that groups separate units into one constituent structure, and **generalization** is the basic merging mechanism that replaces multiple chunks with a generalization. This intuition runs into a limitation when both form and meaning have to be considered. Induction in context free grammars are driven primarily by statistical information (which can be extremely sophisticated), but in construction grammar semantics play a much more foregrounded role. Not only are composition and generalization driven by semantic similarity, but the internal semantic structure of each chunk has to be induced as well. In describing how grammar induction works for a construction grammar, it is necessary to first lay out the elements that the learner considers in its hypothesis space.

3.1 Hypothesis space of construction grammars

Recall that within the framework of construction grammar, both lexical units and phrasal/clausal units are represented as constructions. Each construction can specify a set of constituents in addition to form and meaning relations. Without delving into phonology and morphology, which are concerned with constituency structure among phonological and morphological units, a lexical construction maps an orthographic form to a meaning representation. The meaning of a lexical construction, as demonstrated in the following examples, can range in complexity from a single schema or ontological type (*yao4*, which means medicine, below left) to a set of schemas with role bindings and type restrictions (*chi1*, which means eat, below right). It can be seen that even word learning in this paradigm goes beyond simple one-to-one mappings between symbolic units, and that the formalism is capable of representing layers of meaning that are of different degrees of relevance to a lexical item.

```
Construction YAO4-N
  subcase of Morpheme
  form
    constraints
      self.f.orth <-- "yao4"
  meaning : @Medicine
```

```
Construction CHI1-V
  subcase of Morpheme
  form
    constraints
      self.f.orth <-- "chi1"
  meaning : EAT
    evokes EVENT_STRUCTURE as event_structure
    constraints
      event_structure.inherent_aspect <-->
      self.m.inherent_aspect
```

Figure 3.1 Two lexical items in ECG: *yao4* (medicine) and *chi1* (eat). A lexical construction maps a form to a meaning and has no constituents. The meaning of a lexical construction can be quite simple, such as the ontological type on the left, or very structured, such as the process on the right.

Simple as it seems to represent lexical meaning, it is important to note that a lot of intricacy in the meaning representation is encapsulated in the schemas. As expected, the EAT schema, shown in Figure 3.2, captures the fact that eating is an action involving two participants (the ingester and the ingested) by inheriting these roles from the Ingestion schema. Additionally, it captures the knowledge that the eating process involves two additional sub-processes — chewing and swallowing — by evoking those schemas and setting up bindings between the eater and chewer/swallower and between the eatee and the chewee/swallowee[21].

```
schema EAT
   subcase of INGESTION
      evokes CHEW as chew
      evokes SWALLOW as swallow
      roles
         ingester : @Entity      // inherited
         ingested : @Entity      // inherited
      constraints
         ingester <--> protagonist      // inherited
         inherent_aspect <-- @Inherent_Activity
         ingester <-- @Animate
         ingested <-- @Manipulable_Inanimate_Object
         ingester <--> chew.ingester
         ingested <--> chew.ingested
         ingester <--> swallow.ingester
         ingested <--> swallow.ingested
```

Figure 3.2 The EAT schema captures not only two core participants in the scene but also sub-process relations between eating, chewing and swallowing. The complexity in meaning leads to a large hypothesis space for phrasal and clausal constructions.

Unlike lexical constructions such as YAO4-N and CHI1-V, phrasal constructions are composed of constituents whose form can be ordered and whose meaning can be tapped into. The learned construction, XI1XI1-CHI1-YAO4, taken from the example in Chapter 1 and shown in Figure 3.3, illustrates the hypothesis space that a construction grammar learning model can explore. Intuitively, a number of learning choices are made in this learning scenario. One choice is in the constituency (or branching) structure: the learner can put all three constituents, x0 (XI1XI1), c1 (CHI1) and y2 (YAO4) in a flat structure inside one construction like it has done here, or in a binary branching hierarchy. The current model sees no a prior reason to assume binary branching, but it may still restrict the number of constituents that can be assembled into a construction at any given time to reflect some sort of working memory constraint. A second

[21] The three schemas, Eat, Chew and Swallow as defined all have the inherited role names of ingester and ingested, but for ease of distinction they will be referred to as eater/chewer/swallower and eatee/chewee/swallowee respectively.

choice is made in the ordering constraints amongst the three constituents: the phrase *xi1xi1 chi1 yao4* can be generated from different sets of form constraints, from unordered to fully ordered. Here the choice is to write down the most restrictive pairwise ordering using the **meets** relation, but the less restrictive **before** relation is also consistent with the data. A third choice, contributing the most to the size of the hypothesis space, is in the meaning representation of this construction. In this example, the learner gains three major pieces of information from the situational context: (1) An eating event is about to happen. (2) The eater is the child XiXi. (3) The eatee is the medicine.

```
Construction XI1XI1-CHI1-YAO4
   subcase of CLAUSE
   constructional
      constituents
         x0 : XI1XI1
         c1 : CHI1
         y2 : YAO4
   form
      constraints
         x0.f meets c1.f
         c1.f meets y2.f
   meaning: EAT
      evokes RD as rd0
      evokes RD as rd1
      evokes DISCOURSE_SEGMENT as DS
      constraints
         self.m <--> c1.m
         x0.m <--> c1.m.ingester
         y2.m <--> c1.m.ingested
         rd0 <--> x0.rd
         rd0.referent <--> c1.m.ingester
         rd1.referent <--> c1.m.ingested
         rd1.ontological_category <-- @Medicine
         rd1.discourse_participant_role <-- @Attentional_Focus
         DS.speech_act <-- @Requesting_Action
```

Figure 3.3 Each domain (constructional, form and meaning) contains structures and relations and the current grammar learning problem is a large-scale mapping problem across domains.

However, as in a real learning situation, a great number of other pieces of information are present in the surrounding. Some may be relevant (e.g. that the child is also the addressee in this utterance) and some probably not (e.g. that the medicine is a cherry-flavored syrup). The learner selects the set of meaning constraints using a number of heuristics which will be described in Chapter 4.

Abstractly speaking, then, learning in construction grammar involves a search over mappings across three domains — constituent structures, form relations, and meaning structures — none of which is given a priori and all of which are determined simultaneously. This stands in stark contrast to work in syntactic grammar induction. The majority of automatic grammar induction work are done in formalisms that focus solely on syntax such as Context-Free Grammar (CFG) and Dependency Grammar (Clark, 2001; Klein, 2005), while more recently efforts are seen in inducing grammars with shallow, generally logic-based semantic representations, such as Lexical Functional Grammar (LFG) (Burke, Lam, Cahill, Chan, O'Donovan, Bodomo, Genabith & Way, 2004; Cahill, Burke, O'Donovan, Riezler, van Genabith & Way, 2008), Tree-Adjoining Grammar (TAG) (Chen, Bangalore & Vijay-Shanker, 2005; Xia, Han, Palmer & Joshi, 2001), and Combinatory Categorial Grammar (CCG) (Hockenmaier & Steedman, 2002; Zettlemoyer & Collins, 2007). Even so, semantics is secondary to syntax if at all present and are limited to pre-defined syntax-driven combinatorial rules in these systems.

As some of that work demonstrates, grammar induction is a notoriously difficult machine learning problem even with fairly rigid and shallow semantic representations. In a way, traditional grammar induction tasks are so difficult precisely because the grammars are underdetermined given little or very shallow semantic representations — judging purely from syntax, many parses look just as good as any other. In contrast, the model here is able to rely on the richness of the semantics in order to hone in on the correct constructions. However, the inclusion of unification-based semantics makes for a potentially unbounded search space. By disallowing cycles in the semantic feature structures (as per standard unification grammars) and using a set of predefined meaning schemas with fixed sets of features and discrete fillers, the search space can be made finite, though the number of possible sets of mappings across the three domains is still exponentially big.

A related work on grammar induction in construction grammar (Alishahi, 2008; Alishahi & Stevenson, 2008) sidesteps this issue by treating constructions as probabilistic clusters of verb frames. Two major simplifying assumptions are made in their work: (1) It defines constructions as probabilistic clusters of verb argument frames. Constructions do not introduce novel meaning components and instead only provide a distribution over values of meaning features. The learner problem is reduced to a clustering problem based on six features with fixed values. (2) The learning utterance-scenario pairs are constructed so that the exact meaning of each utterance is supplied in the input, eliminating uncertainty and noise. Furthermore, the learner is equipped with predefined syntactic patterns (e.g. arg1 verb arg2 arg3), dramatically limiting the hypothesis space. While this formulation of the learning problem provides a parsimonious mathematical model of comprehension and production, it suffers from problems of cognitive plausibility. The representation fails to capture the full extent of compositionality in construction grammar, particularly the interaction between verb arguments and constructional arguments which may differ in number and type restrictions. From a learning standpoint, the assumption that the learner begins with pre-formed notions of syntactic pattern is a presumptuous one to make.

It has to be acknowledged, however, that the learning problem is not tractable without some form of learning bias. The current learner, in preserving cognitive plausibility, relies on a set of structural biases in the form of learning operations and representational restriction to limit the search over the hypothesis space. This framework has been initially developed by Chang to study the acquisition of early argument structure constructions in English (Chang, 2008; Chang &

Gurevich, 2004). It is extended here with a structured model of context to cope with context-reliant linguistic phenomena such as argument omission (Chang & Mok, 2006a; b) and a probabilistic version of Embodied Construction Grammar (Bryant, 2008b) is adopted. The current language understanding framework features a heavier reliance on the situational context, which is reflected both by constructions that express contextual restrictions and a context-aware constructional analysis process as described in Chapter 2. In light of these changes and the introduction of a probability model to the grammar formalism, learning operations proposed in (Chang, 2008) are updated and new operations are introduced. The overview of these learning operations is given in this chapter and their technical details are the subjects of the next three chapters.

A bit of terminology is in order here. Recall that the learning model is a comprehension-driven loop which iterates through transcripts of parent-child interaction, analyzes each utterance, attempts to learn from utterances that it does not fully understand, and periodically reorganizes the grammar. We will refer to each of these transcripts as a **dialogue**, and each sets of learning operations between utterances as a **learning episode**. Obviously, learning can take place after both adult and child utterances in a child learner, but since production is not modeled in this dissertation, no learning is performed after a child utterance. For consistency, however, learning episodes are numbered regardless of whether any learning takes place so that the episode count is the same as the total number of utterances encountered by the learner. Finally, in practice, the same set of dialogues may be used for multiple **iterations** to improve learning results.

3.2 Searching in the hypothesis space

Treating each grammar (the collection of schemas and constructions) as a state, the grammar learning procedure can be conceptualized as a search through the state space of possible grammars in ECG. Unfortunately, this space is infinite and there is no hope of arriving at the probabilistically correct grammar using a blind search. The main thrust of this dissertation is twofold: a richly-structured meaning representation in the form of embodied meaning schemas that helps locate the learner in a good region of the hypothesis space, and a set of cognitively motivated learning operations that navigate the learner through it.

At the risk of oversimplification, we appeal to a context-free grammar (CFG) based intuition behind each operation while keeping in mind that the problem at hand is immensely more complex due to the semantic structures. Roughly speaking, concrete constructions correspond to a rule or production, lexical constructions correspond to terminals in CFG, and non-lexical constructions correspond to non-terminals. Abstract constructions have no direct equivalent in CFG: they can often function like unary rewrite rules but they also contain the shared structures amongst its children, so this is where the CFG analogy starts to break down.

The two primary learning operations, composition, and generalization, only add to an existing grammar. The four refinment operations, construction revision, constituent omission, category merge, and category expansion provide ways to fine-tune the grammar and both add and subtract from the grammar (depending on implementation). Finally a catch-all decay operation periodically removes unused constructions from the grammar to keep analysis ambiguity under control. The overview of these operations is described in Figure 3.4 and the technical details are given in Chapter 4 and Chapter 5.

Learning Operation	CFG	Construction Grammar
Composition adds a new concrete construction that puts together other constructions in the grammar.	given collocate(a, b) add c → a b	given a and b used in the same analysis a and b share contextual references add cxn c with constituents a and b
Generalization adds new grammatical categories and generalized concrete constructions based on similarly structured constructions in the grammar.	given a → b c d f → g c d add x → b x → g y → g c d	given cxn a with constituents b, c, and d cxn f with constituents g, c, and d condSub(a, f) OR condSub(f, a) add abstract cxn x with subcases b and g cxn y with constituents x, c, d
Construction revision takes existing constructions in grammar and creates a modified version which is allowed to compete with the original one.	given a → b d collocate(a, c) add a → b c d	given cxn a with constituents b, d cxn f with constituents g, d NOT condSub(a, f) AND ((sem(a, f) AND NOT syn(a,f)) OR (syn(a, f) AND NOT sem(a, f)) collocation of c with a, b or d add cxn a with constituents b, c, d
Constituent omission takes pairs of minimally contrasting constructions and creates a new construction with omissible or optional constituents that can take the place of the pair.	replace a → b d a → b c d with a → b [c] d	given syn(a, f) AND sem(a, f) replace cxn a with constituents b, d cxn f with constituents b, c, d with cxn a with constituents b, [c], d
Category merge takes two or more overlapping constructional categories and smash them together into one category.	replace x → a | b y → a | c with x → a | b | c	given abstract constructions x, y sem(x, y) replace cxn a with parents x, y cxn b with parent x cxn c with parent y with cxn a, b, c each with parent x
Category expansion takes a constructional category and extends it to non-member items.	given x → a distributional sim. between a and b add x → b	given abstract construction x sem(x, b) add cxn b with parent x
Decay removes unused constructions from the grammar	remove a → b c d	remove construction a

Figure 3.4 The learner operations provided in the learning model and their analogue in a CFG-based system. Brackets denote omissible constituents. sem(x, y) denotes semantic subsumption, syn(x,y) denotes syntactic subsumption, and condSub(x, y) denotes conditional subsumption, all of which will be defined in Section 3.4.

While many more operations can be implemented in the model, this set of 7 operations provides a reasonable point of departure for a model of grammar learning. The skeletal structure of the learning model can conceptualized as the simple loop in Figure 1.5, reproduced here:

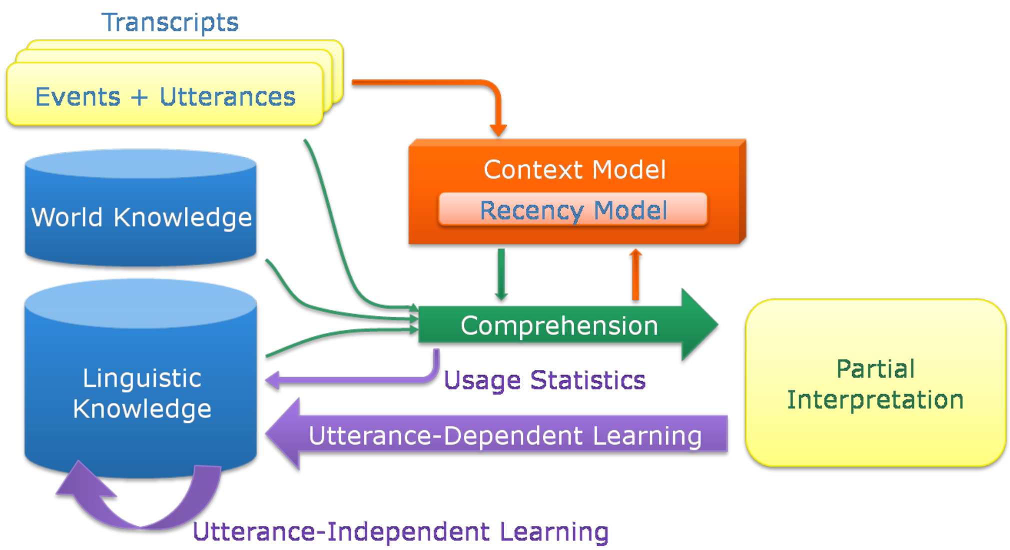

Figure 3.5 The simple comprehenshion-driven learning loop implemented in the model.

3.3 Overview: learning model

The learning operations are grouped here based on whether each derives information from a situated utterance and its analysis. The utterance-dependent learning operations are composition, construction revision, and constituent omission[22]. The utterance-independent operations are generalization, category merge, and category expansion, which do not incorporate any new information from the current analysis. In order to avoid an exhaustive search over the

[22] This may be less obvious from the previous descriptions and the rationale has to do with how one construction may be embedded in another in an analysis. Consider the possibility of a series of operations that first compose XI1XI1 (XiXi) with CHI1 (eat) to get the construction XI1XI1-CHI1 (XiXi eat) and then compose XI1XI1-CHI1 (XiXi eat) with YAO4 (medicine) to get the construction XI1XI1-CHI1-YAO4 (XiXi eat medicine). The fact that both XI1XI1-CHI1 and XI1XI1-CHI1-YAO4 both exist in the grammar does not mean that the YAO4 constituent in XI1XI1-CHI1-YAO4 is omissible precisely because XI1XI1-CHI1-YAO4 takes XI1XI1-CHI1 as a constituent. The learner must wait till it encounters an utterance with only *Xi1+Xi1 chi1*.

entire grammar at every learning episode, the utterance-independent operations are triggered on the constructions that are most recently used (i.e. used in an analysis, newly created, or recently modified).

```
given Grammar g, Dialogues D, MaxIteration n

for i from 1 to n:
  for dialogue d in D:
    for item t in d:
      if (t is an event annotation):
        updateContextModel(t)
      else if (t is an utterance):
        Analysis a = chooseBestAnalysis(t)
        Set<Construction> seeds = getRecentlyUsedCxns(a)
        updateGrammarUsageStatistics(g, a)

        reviseConstructions(g, a)
        checkForOmission(g, a)
        composeConstructions(g, a)

        seeds = generalizeConstructions(g, seeds)
        seeds = mergeCategories(g, seeds)
        seeds = expandCategories(g, seeds)
        decay(g);
```

Figure 3.6 Pseudo-code description of the learning algorithm. Given an initial grammar *g* and a set of dialogues *D*, the learner iterates over the dialogues *n* times, initiating a context model update after each event annotation and a learning episode after each utterance. The learning sequence begins with the analysis of the utterance (including context fitting, which may re-rank the returned analyses), after which usage statistics are gathered based on the top analysis.

3.4 Structural comparison of constructions

As has been discussed all along, a number of learning operations on a construction grammar require structural comparison of constructions. For example, in generalization, the two specific constructions must have compatible constituency structure as well as form and semantic constraints, though they need not be identical. It is therefore important to define construction comparison and compatibility precisely.

The structural comparison procedure cannot be carried out using textual comparison between constraints written in ECG as there are many ways of expressing equivalent constraints. As an example, the set of constraints {a <--> b, b <--> c} is equivalent to {a <--> b, a <--> c} due to the transitivity in unification semantics. A number of different algorithms can be used to assess equivalence; the one adopted by this learning system is described here. Given any two constructions α and β,

- a **constituent mapping** is a one-to-one correspondence between constituents of α and constituents of β. The correspondence between a pair of constituents is defined by a correspondence function f. Depending on the learning operation, this function can be an equality (exact constructional type constraint match), constructional subtype, or semantic subtype match.
- an **evoked role mapping** is a one-to-one correspondence between evoked items of α and evoked items of β. The correspondence function between a pair of evoked items can be an equality (exact semantic type constraint match) or semantic subtype match.
- α **syntactically subsumes** β given a constituent map m if the set of (mapped) form constraints in α are more general than those in β. For example, the **before** relation subsumes the **meets** relation: any pair of constituents that satisfy the **meets** relation also satisfy the **before** relation. To determine if α syntactically subsumes β given a constituent map m:
    - A form constraint matrix is computed for β.
    - Each form constraint from α is mapped to β's term using m and is verified on the constraint matrix.

    - If every such constraint is covered by the matrix, α syntactically subsumes β.
- α **semantically subsumes** β given a constituent map m and an evoked item map e if (1) the meaning pole type of α is a subtype of that of β, and (2) the set of (mapped) meaning constraints in α are more general than those in β. The spirit behind the algorithm to check for semantic subsumption is similar:
    - A feature structure representation is instantiated for the meaning of β (therefore incorporating all of β's semantic constraints).
    - Each meaning constraint from α is mapped to β's term using m and e and is verified on the feature structure.
    - A unification constraint from α is contained in β if both slot chains in the mapped constraint point to the same slot in the feature structure.
    - An assignment constraint from α is contained in β if the referenced slot has the same atomic filler or an equally or more restrictive type constraint as the one specified by the constraint.
    - An assignment constraint from α is relaxable in β if the type constraint of the referenced slot shares a common ancestor with the type specified the constraint.
    - If all constraints from α are contained in β, α semantically subsumes β.
- Construction α **subsumes** construction β if there exist a pair of constituent map and evoked role map $\langle m, e \rangle$ such that α subsumes β both syntactically and semantically.
- Construction α is **equivalent** to construction β if they are mutually subsumed.

- Construction α **conditionally subsumes** construction β if there exist a pair of constituent map and evoked role map <*m*, *e*> such that (1) α subsumes β syntactically, and (2) α subsumes β semantically with up to N relaxable assignment constraints.

Notice that this definition is flexible enough that two constructions with different numbers of constituents can still be structurally aligned. Note also, however, that optional and omissible constituents are not given particular considerations in this definition, in part because omissible constituents are probabilistically defined. This is something that can be improved on as part of future work, but for now, the model uses the stated definitions to discourage structurally aligning constructions with different numbers of optional and omissible constituents. With this, we will launch into the details of the composition and generalization operations.

Chapter 4.

Creating Structure in a Grammar

This chapter details the two basic mechanisms through which the learner creates structure in the grammar: the composition and the generalization operations. Composition creates concrete constructions out of smaller ones by positing form and meaning relations between them, while generalization relaxes constraints on existing constructions and in the process creates linguistic categories. This and the next chapter examine dialogue01 from Chapter 1, where the father offers medicine to the child, to illustrate the various learning operations.

4.1 Composition

The information available to the learner at the start of this learning episode is shown in Figure 4.1. Analyzing with an incomplete grammar (in fact, a lexicon-only grammar), the learner obtains a multi-rooted analysis of the phrase *xi1+xi1 chi1 yao4* (XiXi eat medicine). As a result of the context fitting process, the learner finds shared references between those pieces of analyses. For example, the learner finds that the word *xi1+xi1* refers to the same child (CHI) as the eater of the EAT event denoted by the word *chi1*. It also realizes that *yao4* refers to the same entity (Cough Syrup) as the eatee role of that EAT event. These pairs of shared references are denoted in Figure 4.1 with bold, dashed arrows. With these shared references supplied by context, the learner is able to begin the composition operation.

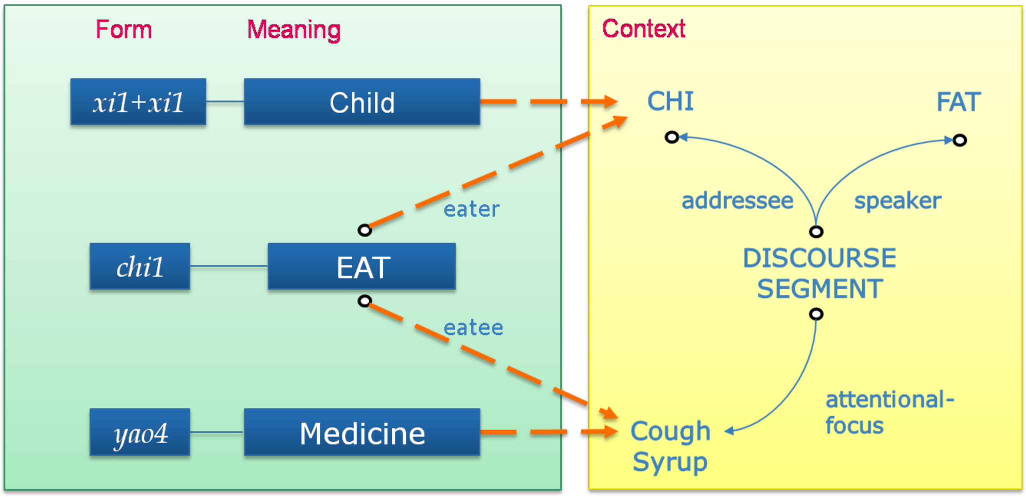

Figure 4.1 Pictorial representation of the resulting analysis from using a lexicon-only grammar to analyze the utterance *xi1+xi1 chi1 yao4* (XiXi eat medicine). On the left, no semantic relations are given in the SemSpec between the lexical constructions due to the lack of phrasal constructions. On the right, the context model contains rich semantic and discourse relations between various entities. After context fitting, the learner finds shared contextual references (bold dashed arrows) which form the basis of the composition learning operation.

Navigating the SemSpec and context fitting output

The clean diagram above belies the complexity of the constructional analysis of a seemingly innocuous utterance such as this. As explained in Chapter 2, the output of the constructional analysis is constructional tree and a semantic specification. Both the constructional tree and the semantic specific can be represented in the same feature structure, shown in Figure 2.2. As was illustrated in the directed acyclic graph representation in Chapter 2, there can be multiple paths that traverse from the ROOT of the analysis to any given slot. This presents a design decision for the learning model: recall that the goal of the composition operation is to specify semantic bindings between slots (e.g. that the eater of the EAT action denoted by *chi1* is the same as the human denoted by *xi1+xi1*). If there are multiple ways to refer to each slot (i.e. multiple essentially equivalent slot chains), the learner could benefit from consistency in its choice when creating a new construction.

A simplified version of the slot chains to relevant semantic slots is given in Figure 4.3. A semantic slot is considered relevant for learning if its type constraint is a subcase of the PROCESS schema or the Entity ontology type. This table shows on the left all possible paths from ROOT to each relevant slot in the three incomplete analyses, leaving out ROOT.s from each chain[23] and redundant DISCOURSE_SEGMENT slots for brevity. The middle column gives the type constraint on each of these slots as well as the fillers, if present. On the right are suggested context elements matched to each slot, given by the context fitter. None of the PROCESSes (Eat, Chew, and Swallow) have been fitted because the utterance is an imperative (i.e. no complete fit) and no similar instances of eating have occurred in context either (i.e. no partial fit).

Five slots are fitted with a context element during context fitting, three of which are assigned to the child and two to the cough syrup. The fact that separate slots are fitted to the same entity in context (or contextually unified) indicates that a semantic unification constraint can be proposed to unify these roles in a new construction. The serial nature of spoken utterances in a word order language such as Mandarin or English yields natural form constraints which can be paired with the semantic constraints. More of the meaning structure in the new construction also has to be determined. While the new construction needs to capture the fact that the protagonist of the EAT event (CHI1-V.m.protagonist) is the same as the reference of the name XI1XI1-N (XI1XI1-N.m), there are other constraints that can be learned, some helpful and some not. For example, the utterance can be taken as evidence that XI1XI1-N can only be used when the corresponding child is also the addressee, but such strict contextual restrictions is unhelpful to the grammar in the long run. On the other hand, in a case of argument omission, the learner may in fact want to learn that the subject can be omitted if it is also the addressee.

[23] The ROOT construction with the s constituent is an artifact of the current analyzer implementation and can be ignored for the purpose of this technical discussion.

```
ROOT
  s:   XI1XI1-N
         m  [1]    @Child
         rd  RD
              referent                [1]
              ontological_category [1]

  DS:  DISCOURSE_SEGMENT
         speaker      [2]    @Human
         addressee  [3]    @Human
         speech_act  SPEACH_ACT
                          speaker       [2]
                          addressee   [3]

  m:  [1]
```

```
ROOT
  s:   CHI1-V
         event_structure:  EVENT_STRUCTURE
                               inherent_aspect:  @Inherent_Activity

         m:    [4]  EAT
                      protagonist:    [5]    @Animate
                      ingester:         [5]
                      ingested:         [6]    @Manipulable_Inanimate_Object
                      instrument:               @Entity

                      chew:      CHEW
                                   inherent_aspect:  @Inherent_Activity
                                   protagonist:     [5]
                                   ingester:          [5]
                                   ingested:          [6]

                      swallow:  SWALLOW
                                   inherent_aspect:  @Inherent_Achievement
                                   protagonist:     [5]
                                   ingester:          [5]
                                   ingested:          [6]

                      inherent_aspect:  @Inherent_Activity
  m:  [4]

  DS:  DISCOURSE_SEGMENT
              speaker      [7]    @Human
              addressee  [8]    @Human
              speech_act  SPEACH_ACT
                               speaker       [7]
                               addressee   [8]
```

```
ROOT
  s:   YAO4-N
         m  [9]    @Medicine

  DS:  DISCOURSE_SEGMENT
         speaker      [10]   @Human
         addressee  [11]   @Human
         speech_act  SPEACH_ACT
                          speaker       [10]
                          addressee   [11]

  m:  [9]
```

Figure 4.2 Feature structure representation of the same constructional analysis of X*i1+xi chi1 yao4* prior to context fitting. There are three ROOTs in this analysis, one for each word. Each root is given its own DISCOURSE_SEGMENT schema because they may come from different parts of the utterance and carry different speech acts. The DISCOURSE_SEGMENTs are unified in the context fitting process, yielding the (simplified) co-indexed slot chain representation shown in the next figure.

Co-indexed Slot Chains		Slot Filler	Context Element
DS.speaker DS.speech_act.speaker	[2]	(unfilled) *@Human*	FAT
DS.addressee DS.speech_act.addressee	[3]	(unfilled) *@Human*	**CHI**
XI1XI1-N.m XI1XI1-N.rd.referent XI1XI1-N.rd.ontological_category	[1]	(unfilled) *@Child*	**CHI**
CHI1-V.m		EAT	null
CHI1-V.m.chew		CHEW	null
CHI1-V.m.swallow		SWALLOW	null
CHI1-V.m.protagonist CHI1-V.m.ingester CHI1-V.m.chew.protagonist CHI1-V.m.chew.ingester CHI1-V.m.swallow.protagonist CHI1-V.m.swallow.ingester	[5]	(unfilled) *@Animate*	**CHI**
CHI1-V.m.ingested CHI1-V.m.chew.ingested CHI1-V.m.swallow.ingested	[6]	(unfilled) *@Manipulable_ Inanimate_Object*	Cough Syrup
CHI1-V.m.instrument		(unfilled) *@Entity*	null
YAO4-N.m	[9]	(unfilled) *@Medicine*	Cough Syrup

Figure 4.3 The leftmost column shows the set of (simplified) slot chains from each ROOT to each slot that is relevant to the learner (any slot whose type constraint is a subcase of PROCESS or @Entity). The middle column shows the slot filler as returned by the analyzer (either structures or type constraints on unfilled slots), and the matching context element suggested by the context fitter. Redundant DISCOURSE_SEGMENT schemas are removed for space.

In sum, for each new construction, the learner needs to determine:

- the appropriate constructions to recruit as proposed constructional constituents
- the order constraints between the new constituents
- the relevant contextual unifications for the new construction
- a coherent overall meaning for the new construction

Associative learning of constituents and meaning constraints

To concretize the task , the result of the composition operation is a set of candidates each consisting of the following:

- *cxnName*: a name for the new construction
- { c : constructional constituent specified by type and local name }
- { p : constructional parents of the new construction }
- { o : ordering constraints in the form block }
- *mType* : type of the meaning pole
- { e : evoked role in the meaning pole specified by type and local name }
- { u : unification constraints in the meaning pole }
- { a : assignment constraints in the meaning pole }

The current learning algorithm relies on a form of associative learning that is very structured while care is taken to reduce noise. The use of contextual information to guide utterance interpretation greatly focuses the analyses to a few nearly correct ones and amounts to the use of attentional and intention cues by young children. Even then, with a multi-word utterance by the parent, multiple perspectives and levels of granularity of understanding an event, as well as possibly multiple relevant events in a scene, the scope for a new construction search is big. In an extreme approach, the learner can memorize the entire constructional forest and semantic specification as a new construction, set the generalization mechanism to work and hope for the best. While this approach may be computationally feasible given copious amounts of data, its cognitive plausibility is challenged by both its memory requirements and psychological evidence about how children learn.

The learning model therefore adopts a moderated associative learning approach which selectively associates meaning relations with form relations using a set of heuristics. For simplicity, the composition procedure first searches for single-unifier compositions, i.e. forming a new constructions based on constructions in the analysis that are semantically connected through one single context element. In the example, XI1XI1-N and CHI1-V are connected through the child CHI, and CHI-V and YAO4 are connected through the Cough Syrup. A single-unifier composition puts either XI1XI1-N and CHI1-V together or CHI-V and YAO4 together but not all three.

Once the learner locates all the single-unifier compositions, it can easily group the semantically connected ones into a larger construction by combining the respective form and meaning constraints as long as the constituents are (nearly) adjacent. In the case of the example, XI1XI1-N, CHI1-V, and YAO4 are semantically connected through the EAT schema, so a construction with all three as constituents can be learned by the learner. The formation of single-unifier compositions is most illustrative of the composition operation and the rest of this section will focus on this topic.

Constructional constituents and ordering constraints

In choosing single-unifier compositions, the learner looks for two or more slots in the SemSpec that share references to the same context element. Each of these slots can be traced back to a (set of) constructions, which can be in one of the configurations shown in Figure 4.4. Each configuration shows the input string '*a x b x y*' on top, the constructional tree in the top half, and the semantic DAG as well as contextual references in the bottom half. In this notation, each construction covers the alphabets in its name, e.g. CXN-A is a lexical construction with an orthography of 'a', and m denotes the respective meaning pole schemas.

In the simplest case such as (a), CXN-A, CXN-X and CXN-B share contextual references to CE1 and are selected as constituents of a new construction and obey the form constraints of CXN-A meets CXN-X and CXN-X meets CXN-B. These form constraints are deduced from the spans of each of these constructions in the input string: constituents are ordered based on their spans[24] and a form constraint is written between each constituent and its succeeding one. The most restrictive form constraints is always used, so adjacent constituents are assigned a meets constraint and non-adjacent constituents are assigned a before constraint.

As the learner builds up phrasal constructions, the constructional trees will gain depth in their derivational structures. Suppose that construction CXN-AB, which covers CXN-A and CXN-B as long as CXN-A comes before CXN-B, has been learned. Configurations such as (b) and (c) are now possible. In these and all other cases, the constructions closest to roots, i.e. the longest spanning constructions, are chosen as constituents so long as their spans do not overlap. In (b), CXN-AB spans [0, 3] and CXN-Y spans [4, 5], and they are chosen as new constituents with the expected ordering constraints. (c) is possible, however, due to the robust analyzer's ability to skip over an unconnected root within an analysis (refer to Section 2.3 for explanation of the analyzer). In this case, CXN-AB's [0, 3] span engulfs CXN-X's [1, 2] span. The construction resulting from composing CXN-AB with CXN-X does not have acceptable form constraints because the current ECG implementation disallows crossing constituent spans. To remedy this situation, CXN-AB is "flattened" inside the new constructions, i.e. the constituents of CXN-AB, CXN-A and CXN-B, are used as constituents along with CXN-X. The learned construction will have the same constituent and form structure as in configuration (a), and the meaning constraints are adjusted accordingly.

---

[24] Spans are given in standard NLP notation as denoted by the subscripts for the string *a x b x y*:

$_0 a_1 x_2 b_3 x_4 y_5$

CXN-A, which covers 'a', thus has a span of [0, 1].

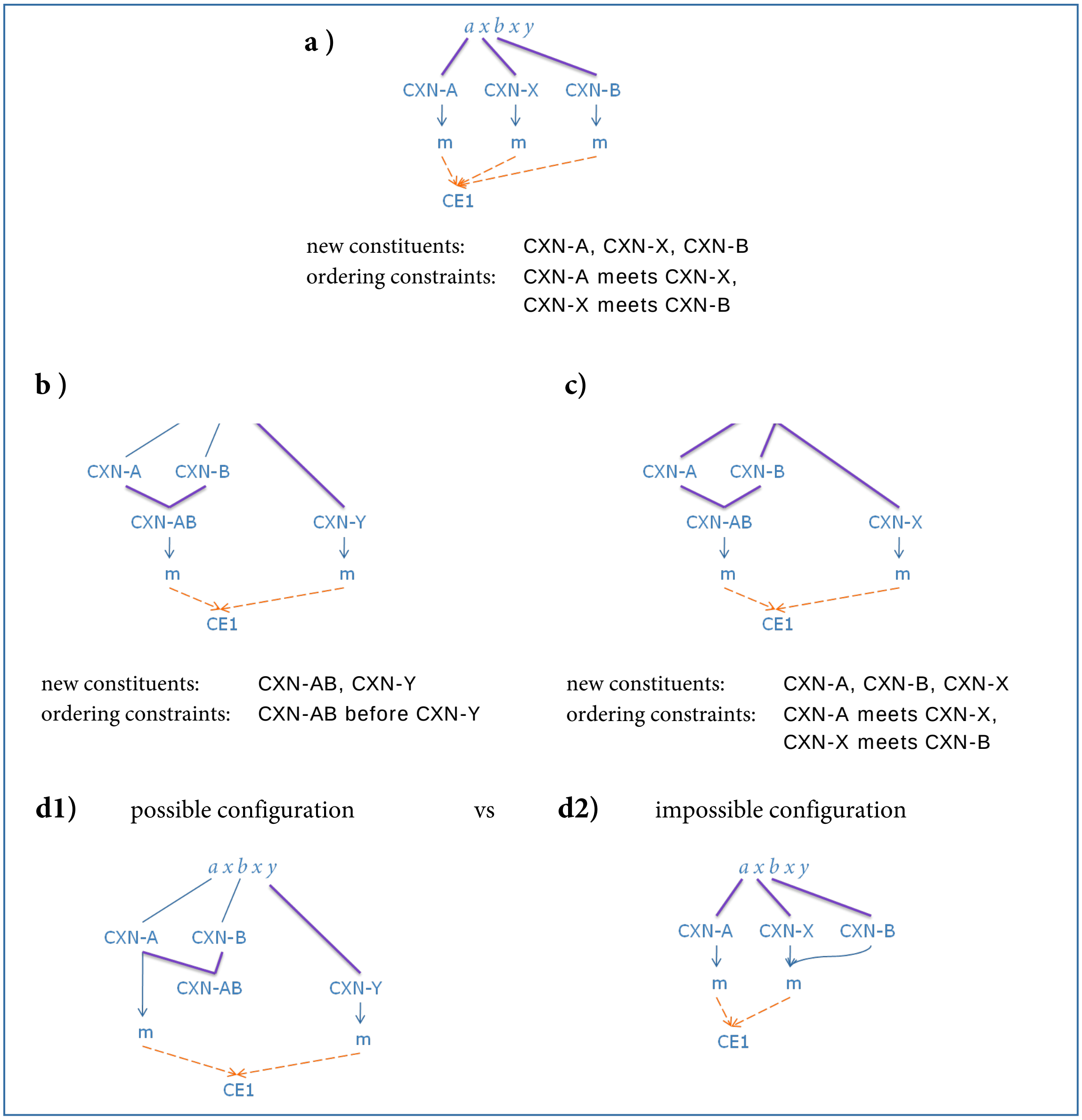

Figure 4.4 The constituency structure of a new construction depends on the configuration of constructions used in an analysis. (a) and (b) are straightforward cases where the topmost (largest spanning) constructions are selected as new constituents. (c) arises due to the analyzer's robust parsing ability which skips over a completed root inside another root. In this case the spans of the topmost constructions cross (which leads to a malformed ECG construction) and the learner must "flatten" the topmost constructions in order to use them as constituents of a new construction.

Even though more than one construction can be on the paths to a slot, those constructions are guaranteed to be in the same constructional derivation (d1) rather than under separate roots (d2). (d2) is not achievable because the constraint that co-indexes `CXN-X.m` with

CXN-B.m must by definition reside in a third construction that takes CXN-X and CXN-B as constituents. For the coindexation to be present in the SemSpec this third construction will have been introduced into the analysis. (d1), on the other hand, is commonly encountered, and by the constituent selection heuristic described in the last paragraph, CXN-AB and CXN-Y are chosen as new constituents. The resulting construction from (d) has the same constituent structure as that of (c), but with different semantic constraints, obviously.

Constructional parent

New constructions need to be assigned constructional parents for practical reasons[25], and it was mentioned in Chapter 1 that the learner begins with vague notions of word-like things, phrase-like things, and sentence-like things. These are given in the initial grammar as the abstract constructions MORPHEME, WORD, PHRASE, and CLAUSE. New compositions are either PHRASEs or CLAUSEs, depending on their meaning pole type. The constructions that have an overall meaning of a process are made subcases of CLAUSE, and those that do not are made subcases of PHRASE. How the construction meaning is determined is our next topic of discussion.

Meaning pole type and meaning constraints

At the same time as selecting the constructional constituents, all paths to each reference-sharing slot can be computed, some of which are then used to create new unification constraints in the new construction. Taking Figure 4.3 again as an example, three slots share references to CHI and two are "rooted" by a construction. Once those two constructions, XI1XI1-N and CH1-V, are chosen to be constituents of a new construction, the meaning pole can be decided. For purely aesthetic reasons, easy-to-understand constraints such as XI1XI1-N.m <--> CHI1-V.m.ingester are preferred over the equivalent XI1XI1-N.rd.referent <--> CHI1-V.m.swallow.ingester.

[25] The analyzer needs to know which constructions can act as a ROOT of a constructional analysis.

However, most slot chains to either slot suffice because they are all co-indexed by definition of the two lexical constructions, as long as care is taken in other learning operations to perform structural comparisons between constructions (as discussed in Chapter 3) rather than direct slot chain comparisons .[26] In practice the shortest chain (in terms of number of slots in the slot chain) is chosen for each slot, with a preference for locally defined roles wherever possible.

In addition to selecting unification constraints for the new construction, the learner also attempts to find a coherent meaning pole type, which is particularly important considering the potentially complex network of meaning schemas introduced by the constituents. A few things are desired for the meaning pole: we would like the learner to select an process meaning such as EAT for clausal constructions such as XI1XI1-CHI1-YAO4, and we also need the learner to consider other meaning types such as complex events or image schemas in other constructions. The learner determines the meaning pole type by determining the number of meaning roots and process schemas present. Meaning poles of constituents are meaning roots only if they are not unified with a role in another schema. For example, in XI1XI1-CHI1-YAO4, only the EAT schema is a meaning root since both @Child and @Medicine fill roles of the EAT schema, even though they are all direct meaning poles of the three constituents. On the other hand, in a serial verb construction such as XI1XI1-LAI2-CHI1 (XiXi-Come-Eat), both COME and EAT are meaning roots. The heuristics showin in Figure 4.5 are used by the learner to choose the meaning pole type.

---

[26] There is one additional caveat: slot chains found in the semspec are not guaranteed to be well-formed when used as a meaning constraint in ECG. An untyped role, for example, may be type restricted in a semspec due to its unification with another role and then filled with a structured filler. As a result, there is a path in the SemSpec DAG that goes from the untyped role, `r1`, to a role in the structured filler, `r2`. However, the slot chain `r1.r2` is malformed given the current ECG rules enforced by the analyzer because r1 by itself does not have the necessary type constraint.

	No meaning roots	One meaning root	Multiple meaning roots
No process	any present meaning	meaning root	non-compositional
One process	process	process; evokes meaning root if different	process + evokes meaning roots
Multiple processes	non-compositional	meaning root	non-compositional

Figure 4.5 Choosing a meaning pole based on the number of process and root meanings.

Non-compositional meaning refers to meaning components introduced into the new construction that are not found in any of its constituents. In the case of the serial verb example XI1XI1-LAI2-CHI1 (XiXi-Come-Eat), since both COME and EAT are processes, a reasonable meaning pole for the construction is a SERIAL_PROCESS where COME is its first event and EAT is its second event. Since this meaning is introduced by the new construction, an additional search over appropriate schemas must be performed when learning such a construction.

The current learner contains a proof-of-concept implementation for learning constructions with non-compositional meaning. It searches over its known schemas and looks for schemas that have type-appropriate roles for the multiple meaning roots or processes. Once a list of potential candidates is generated, the semantic appropriateness can be evaluated using the semantic model described in Section 2.3. This learning process is obviously noise-prone. The future work chapter gives more details on how this work can be improved and extended to cover other types of non-compositional meanings and their radial extensions.

Contextual constraints on core roles and speech acts

Finally but not the least importantly, information about the current scene is learned as contextual restrictions on the new construction. While current scene information can be unbounded — it may, for example, include the history of events leading up to the present utterance — and is modulated by a combination of factors such as memory, perceptual and emotional salience in a child learner, the present model limits the learnable constraints to information about the current discourse segment. This includes identities of the participating context elements and information about the current speech act (speaker/addressee information, attentional focus, and speech act type), which are learned as constraints on core process roles.

Operationally, these contextual constraints are captured in constructions as restrictions on RDs, or referent descriptors, also introduced in Chapter 2. In new constructions that have process meaning, one RD is evoked for each core role and two pieces of information are kept: the type constraint and discourse participant role of its referent. In the XI1XI1-CHI1-YAO4 example, an RD (rd1) is evoked for the ingested role and the learner posits that the ingested must be of type medicine as well as be the attentional focus. This is captured with these two constraints:

```
rd1.ontological_category <-- @Medicine
rd1.discourse_role <-- @Attentional_Focus
```

To avoid excessive number of dangling RDs, care must be taken to unify existing ones. In the example, the construction XI1XI1-N already evokes an RD, so the new construction simply evokes an RD (rd0) for its ingester role and unifies it with the RD evoked by XI1XI1-N. Over time, through generalization, these learned constraints on the RDs are relaxed or dropped altogether, but they give a means of expressing contextual constraints on referents when constructions with omissible elements are learned.

Summary: the composition operation

The composition operation, as the primary means of generating new grammatical structure, is driven by contextual unification of meaning slots in a multi-rooted constructional tree. When two slots descending from different roots (e.g. the agent role of an action and the meaning of a proper noun) point to the same entity in context, a new construction is created with a meaning constraint to capture this unification. Ordering constraints are created based on word order in the input string and the constituent structure is derived from the constructional analysis. The meaning pole of the new construction is determined by the combination of process and root meanings brought in by the constituents, which in turns determines the constructional parent of the new construction. A number of semantically-connected single-unifier composition can be combined to create constructions with more constituents and more complex meaning, if desired, as exemplified by the learned construction XI1XI1-CHI1-YAO4 (XiXi eat medicine) in Figure 4.6. If the resulting composition is not subsumed by other constructions in the current grammar[27], the composition is added to the grammar for use in the next cycle of learning episodes.

```
Construction XI1XI1-CHI1-YAO4
    subcase of CLAUSE
    constructional
        constituents
            x0 : XI1XI1-N
            c1 : CHI1-V
            y2 : YAO4-N
    form
        constraints
            x0.f meets c1.f
            c1.f meets y2.f
    meaning : EAT
        evokes RD as rd0
        evokes RD as rd1
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> c1.m
            x0.m <--> c1.m.ingester
            y2.m <--> c1.m.ingested
            rd0 <--> w0.rd
            rd0.referent <--> c1.m.ingester
            rd1.ontological_category <-- @Human
            rd1.discourse_role <-- @Addressee
            rd1.referent <--> c1.m.ingested
            rd1.ontological_category <-- @Medicine
            rd1.discourse_role <-- @Attentional_Focus
            DS.speech_act <-- REQUESTING_ACTION
```

Figure 4.6 XI1XI1-CHI1-YAO4 **(XiXi eat medicine): the learned construction from composing** XI1XI1-N**,** CHI1-V**, and** YAO4-V **based on the input utterance *xi1+xi1 chi1 yao4* in context. This is obtained by combining two single-unifier compositions,** XI1XI1-CHI1 **(XiXi eat) and** CHI1-YAO4 **(eat medicine).**

[27] which occasionally happens when the larger construction exists but the analyzer decides not to use it in the best analysis due to associated semantic or constructional costs.

Here are some other examples of constructions learned through composition in the model:

```
Construction CHE1_HUAI4-c023
    subcase of CLAUSE
    constructional
        constituents
            c0 : CHE1-N
            h1 : HUAI4-V
    form
        constraints
            c0.f meets h1.f
    meaning : BROKEN
        evokes RD as rd0
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> h1.m
            rd0.referent <--> h1.m.protagonist
            c0.m <--> rd0.referent
            rd0.ontological_category <-- @Car
            rd0.discourse_role <-- @Attentional_Focus
            DS.speech_act <-- EXPLAINING
```

Figure 4.7 CHE1-HUAI4 **(car broken) : an early instance of the subject-verb construction that denotes object-state. There is one core role, the** protagonist**, and thus one evoked** RD**; it is filled by the meaning of** CHE1-N**. The car happens to be the attentional focus during the utterance, whose speech-act is explaining.**

```
Construction TI1-c143
    subcase of CLAUSE
    constructional
        constituents
            t0 : TI1-V
    meaning : KICK
        evokes RD as rd0
        evokes RD as rd1
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> t0.m
            rd0.referent <--> t0.m.force_recipient
            rd1.referent <--> t0.m.agent
            rd0.ontological_category <-- @Ball
            rd0.discourse_role <-- @Attentional_Focus
            rd1.ontological_category <-- @Child
            rd1.discourse_role <-- @Addressee
            DS.speech_act <-- APPROVING
```

Figure 4.8 TI1 **(kick): a construction learned when the mother encourages the child to keep kicking the ball. No arguments were expressed in the utterance but the construction keeps two** RDs **for the two core roles, agent, and** force_recipient **which turn out to be the** Addressee **and** Attentional _Focus**, respectively.**

```
Construction NI3-QIAO2-MO3-cN010
    subcase of CLAUSE
    constructional
        constituents
            n0 : NI3_VARIANT-N
            q1 : QIAO2-V
            m2 : MO3-V
    form
        constraints
            n0.f meets q1.f
            q1.f meets m2.f
    meaning : COMPLEX_PROCESS
        evokes RD as rd0
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m.process1 <--> q1.m
            self.m.process2 <--> m2.m
            q1.m.perceiver <--> m2.m.force_supplier
            q1.m.perceiver <--> n0.m
            rd0.referent <--> self.m.protagonist
            DS.speech_act <-- EXPLAINING
```

Figure 4.9 NI3-QIAO2-MO3 **(you see apply): an erroneous (but sensible) construction learned when the mother scolded the child for applying too much lotion all over her hands. The utterance was *ni3 qiao2 mo3 zhe yi1 shou3* (you see apply DUR one hand / you see how you got it all over your hands). The context fitter did not manage to associate the applying event with the percept, but at least found that both the perceiver and the applier are the same person. A** COMPLEX_PROCESS **is proposed as the non-compositional meaning to tie the two processes together.**

```
Construction ZHEI4-SHEN2ME-c157
    subcase of CLAUSE
    constructional
        constituents
            z0 : zhe4_variant-D
            s1 : shen2me-WH
    form
        constraints
            z0.f meets s1.f
    meaning: @Element
        evokes Discourse_Segment as DS
        constraints
            z0.m <--> s1.m
            self.m <--> z0.m
            DS.speech_act <-- REQUESTING_ANSWER
```

Figure 4.10 ZHEI4-SHEN2ME **(this what): A phrase learned when the mother asked the child what an object is. Often *zhei4 shen2me* is the entire utterance for asking "what is this", with the copula so unstressed as to being omitted altogether. Since there is no process in the meaning, the construction is the subcase of a** PHRASE**. However, it still captures the fact that the speech act is a request for answers.**

4.2 Generalization

While the last section gives one basic way to create new concrete constructions, a child's ability to generalize grammatical patterns is vital to the formation of a productive, adult-like grammar. One generalization phenomenon of primary interest is knowledge about argument structures, including, in the case of pro-drop languages, allowable patterns of argument omission. This kind of generalization allows the learner to obtain from *xi1+xi1 chi1 yao4* (XiXi eat medicine) and *xi1+xi1 chi1 fan4 (*XiXi eat rice*)* not only a ingester-EAT-ingested construction, but eventually the more general active transitive construction as well.

Representational choices in generalization

It is important to recognize that there are two operational components in service of generalization in a grammar: the formation of grammatical categories and the replacement of specific constituents in a construction with placeholders that have more relaxed type constraints.

The first operational component, the formation of a grammatical category, is implemented as the creation of an abstract construction in ECG. An abstract construction allows its subcases to go in places wherever something of that abstract type is required. For example, verb phrase (VP) is a grammatical category that encompasses many different types of verb phrases which may not share a lot in common syntactically or semantically except for perhaps a main verb and a process meaning. However, any construction that is a subtype of VP can be joined with the subject in a SUBJECT-VP construction.

The second operational component, the replacement of specific constituents in a construction with placeholders, is implemented as the creation of a new concrete (albeit more general) construction in ECG. For example, a particular kind of verb phrase, such as a Caused-

Motion-With-PP construction, contains constituents that allow different verbs, objects, and prepositional phrases. The Caused-Motion-With-PP construction thus represents a generalization over many verb-island constructions that otherwise need to be enumerated.

Let us examine first the basic case of generalizing across two constructions with equal number of constituents, e.g. WO3-CHI1 (I eat) and NI3-HE1 (you drink), shown below. At the end of generalization, two new categories CAT001:{WO3-N, NI3-N} and CAT002:{CHI1-V, HE1-V} as well as one new concrete construction CAT001-CAT002 are created, as demonstrated in the next set of figures.

```
Construction WO3-CHI1
    subcase of CLAUSE
    constructional
        constituents
            w0 : WO3-N
            c1 : CHI1-V
    form
        constraints
            w0.f meets c1.f
    meaning : EAT
        evokes RD as rd0
        evokes RD as rd1
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> c1.m
            w0.m <--> c1.m.ingester
            rd0 <--> w0.rd
            rd0.referent <--> c1.m.ingester
            rd1.referent <--> c1.m.ingested
            rd1.ontological_category <-- @Rice
            rd1.discourse_role <--
                @Attentional_Focus
            DS.speech_act <-- EXPLAINING
```

```
Construction NI3-HE1
    subcase of CLAUSE
    constructional
        constituents
            n0 : NI3-N
            h1 : HE1-V
    form
        constraints
            n0.f meets h1.f
    meaning : DRINK
        evokes RD as rd0
        evokes RD as rd1
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> h1.m
            n0.m <--> h1.Twom.ingester
            rd0 <--> n0.rd
            rd0.referent <--> h1.m.ingester
            rd1.referent <--> h1.m.ingested
            rd1.ontological_category <-- @Soup
            rd1.discourse_role <--
                @Attentional_Focus
            DS.speech_act <-- REQUESTING_ACTION
```

Figure 4.11 Two constructions to be generalized: WO3-CHI1 **(I eat) and** NI3-HE1 **(you drink) .**

The new category CAT001 contains the shared evoked roles and constraints between WO3-N (I) and NI3-N (you), whereas the new category CAT002 contains shared structures between CHI1-V (eat) and HE1-V (drink) and has an appropriate meaning of INGESTION. These two new categories are used in a new concrete construction as shown in Figure 4.12.

```
abstract construction CAT001
    subcase of Morpheme
    meaning : @Human
    evokes DISCOURSE_SEGMENT as DS
    evokes RD@SCHEMA as rd
    constraints
        rd.referent <--> self.m
        rd.ontological_category <--> self.m
```

```
abstract construction CAT002
    subcase of Morpheme
    meaning : INGESTION
    evokes EVENT_STRUCTURE as event_structure
    constraints
        event_structure.inherent_aspect <-->
            self.m.inherent_aspect
```

```
construction WO3-N
    subcase of CAT001 Morpheme
    meaning
    constraints
        self.m <--> DS.speaker
        rd.discourse_role <-- @Speaker
```

```
construction CHI1-V
    subcase of CAT002 Morpheme
    form
        constraints
            self.f.orth <-- "chi1"
    meaning : EAT
```

```
construction NI3-N
    subcase of CAT001 Morpheme
    meaning
    constraints
        self.m <--> DS.addressee
        rd.discourse_role <-- @Addressee
```

```
construction HE1-V
    subcase of CAT002 Morpheme
    form
        constraints
            self.f.orth <-- "chi1"
    meaning : DRINK
```

```
Construction CAT001-CAT002
    subcase of CLAUSE
    constructional
        constituents
            c0 : CAT001
            c1 : CAT002
    form
        constraints
            c0.f meets c1.f
    meaning : INGESTION
        evokes RD as rd0
        evokes RD as rd1
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> c1.m
            c0.m <--> c1.m.ingester
            rd0 <--> c0.rd
            rd0.referent <--> c1.m.ingester
            rd1.referent <--> c1.m.ingested
            rd1.ontological_category <-- @Food
            rd1.discourse_role <-- @Attentional_Focus
```

Figure 4.12 By generalizing over WO3-CHI1 **and** NI3-HE1**, two abstract constructions —** CAT001 **and** CAT002 **— and a concrete construction that uses those two as constituents are created.**

The rest of this section gives the implementation overview of how candidate constructions are selected and how the new constructions are created. The overall generalization procedure is a loop that is seeded with a set of recently used constructions. Generalization on the seeds is performed as long as they are available, and the new grammar at the end of the process is compared with the old grammar. If the new grammar is preferable to the old grammar, it is adopted by the learner. Preferences of grammars can be expressed in terms of a wide variety of measures, including most obviously Bayesian models that include the data likelihood and probabilistic priors on the kinds of grammars, or an information-theoretic approach such as Minimum Description Length. Here it is unclear whether any of these measures is appropriate since the specific constructions remain in the grammar alongside the generalizations. This issue will be discussed in detail in the final chapter. For the basic model, the new, more general grammar is always accepted.

```
given Grammar g, Set<Construction> seeds

g' = g.makeCopy();

while (indexer has next candidate construction pair {c1, c2}):
  if (alignsStructurally(c1, c2)):
    generalize(g', c1, c2)
  else:
    notate for future refinement

if (acceptChanges(g', g)):
  g = g'
```

Figure 4.13 At each learning episode the learner starts the generalization process using the recently used constructions and tries to make all the generalizations warranted by the data.

Searching for constructions to generalize over

One of the more fundamental questions about language learning is why generalizations occur in the order they do. Two competing hypotheses are explored in the model:

- **Constituent-based generalization**. Constructions are selected for generalization based on shared constituents. XI1XI1-CHI1-YAO4 (XiXi eat medicine) and CHI1-FAN4 (eat rice) may be selected for generalization based on the shared constituent CHI1 (eat), but so may XI1XI1-CHI1-YAO4 (XiXi eat medicine) and XI1XI1-LAI2 (XiXi come) based on the shared constituent XI1XI1-N.

- **Semantic-based generalization**. Constructions are selected for generalization based on meaning similarity as dictated by the schema hierarchy. XI1XI1-CHI1-YAO4 (XiXi eat medicine) and CHI1-FAN4 (eat rice) may be selected for generalization based on the shared meaning of a eating scene. XI1XI1-CHI1-YAO4 has similar enough meaning to NI3-HE1 (you drink) to be generalized, but not with XI1XI1-LAI2 (XiXi come).

After brief experimentation, the constituent-based generalization strategy is found to lead to hasty generalizations across verb scene types. It has been explained that XI1XI1-CHI1-YAO4 (XiXi eat medicine) and XI1XI1-LAI2 (XiXi come) are immediately generalizable using this strategy, leading to a XI1XI1-VP-like construction that has vague semantics for the VP. Similar problems can occur with pairs such as CHI1-FAN4 (eat rice) and CHENG2-FAN4 (load rice). This generalization creates a category of verbs that encompasses CHI1 and CHENG2, something children are not observed to do until much later in the language learning process (Abbot-Smith, Lieven & Tomasello, 2004; Akhtar & Tomasello, 1997; Tomasello, 2003).

The model implemented and described here thus follows the second strategy, which selects constructions that are semantically compatible with the seeds for generalization. The schema lattice acts as a crucial moderator of this generalization process, gradually allowing more semantically divergent constructions to be generalized. Given a seed construction, other constructions with meaning that are its parents, children or siblings are retrieved. This particular requirement is a parameter of the learner which, along with the schema lattice, can be experimented with.

Figure 4.14 shows a small section of the schema lattice in use by the system, where the non-leaf nodes (unshaded) represent the structure of the semantic hierarchy and the leaf nodes are schemas that correspond to meanings of lexical items. For example, within the section of the lattice shown below, a construction with a Translational_Forceful_Motion meaning triggers retrieval of constructions with Forceful_Motion, Translational_Motion, Load, and Translational_Self_Motion. Only the handful of retrieved constructions which align structurally with the seed construction can be generalized; this is the subject of the next subsection.

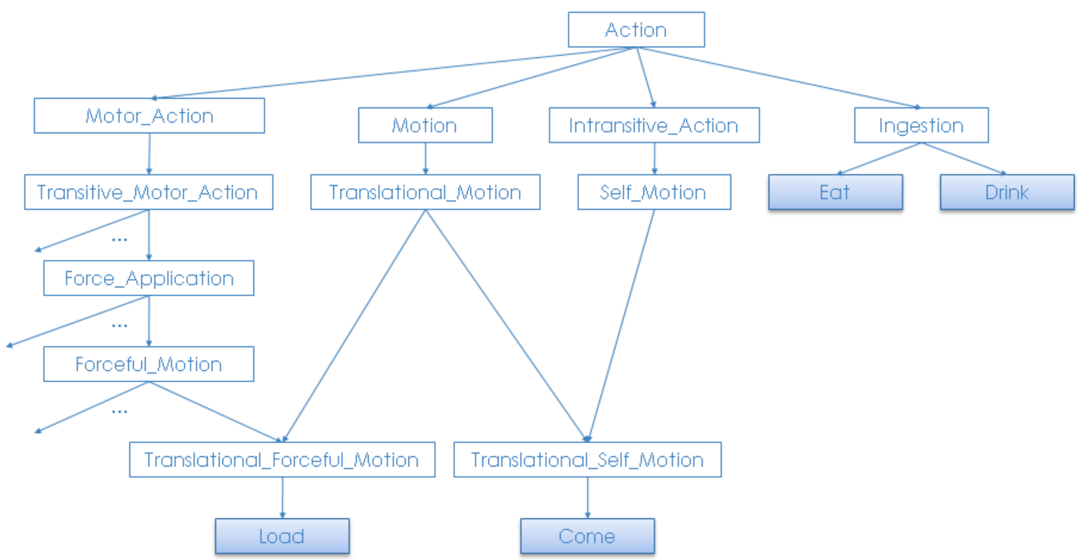

Figure 4.14 A fragment of the process schema lattice used by the learner. The learner requires that only semantically close constructions can be generalized (i.e. parent, children and sibling types). In doing so, this (or similar) schema lattice moderates the pace of generalization.

Structural alignment between candidate constructions

The goal of structurally aligning a pair of candidate constructions is to make sure that they propose comparable meaning relations between their corresponding constituents. Take as an example two constructions such as WO3-CHI1 (I eat) and YAO4-CHI1 (medicine eat), which may be learned from an utterance such as *ba3 yao4 chi1 le* ($\mathrm{CV_{obj}}$ medicine eat ASP). The former posits that the meaning of first constituent is connected to the eater role whereas the latter posits that the meaning of the first constituent is connected to the eatee role. The learner needs to prevent these two constructions from being erroneously generalized over.

Section 3.4 gave the details on how two constructions can be structurally compared. For the purpose of generalization across two constructions, the learner requires that one of them conditionally subsumes the other[28], allowing a small number (3 or 4) of assignment constraints to be relaxed. These assignment constraints, as explained in Section 1.1, express contextual restrictions on core schema role fillers and necessarily differ across situations.

If one construction does conditionally or unconditionally subsume the other, the pair of constructions can proceed to be generalized. Otherwise they are examined for potential revisions. The exact details of revision triggers and revision procedures are given in the next chapter, but the intuition can be taken from the just-mentioned example of WO3-CHI1 (I eat) and YAO4-CHI1 (medicine eat). Since the two forms are similar but they mean different things, by the Gricean assumption (Grice, 1975), there ought to be other differences in the constructions that went unnoticed. The structural alignment process conveniently locates these contrasting pairs as a by-product of searching for possible generalizations and notates them for future examination.

Recursive generalization of constituents

When two constructions are selected for generalization, generalizations of their constituents are also triggered. This is demonstrated in the example of the creation of the lexical categories CAT001:{WO3, NI3} and CAT002:{CHI1, HE1}. This is modeled as a recursive process which triggers the creation of new categories and new concrete constructions.

[28] A more prudent learner can also require that the constructions are equivalent given a constituent map. Due to the limited learning data available, a less exact matching algorithm is adopted to encourage generalization.

The base case of recursion is when either of the constructions is lexical. A new abstract category is created since there is no internal structure to support further decomposition. Otherwise one or two new concrete constructions are returned at the end of the generalization process depending on whether the two specific constructions are of the same length. Pairs of structurally aligned constituents are examined and new constituent types are determined recursively as stated in the pseudo-code in Figure 4.15. Categories are automatically merged if a new category is created over one or more existing categories. The merging operation is further described in Section 5.4.

```
generalize(grammar g', cxn c1, cxn c2, constituentMap m,
evokedRoleMap e)


if (isLexical(c1) || isLexical(c2)):

  return createNewCategory(c1, c2)

else:

  // determine constructional types of new constituents

  for (constituent pair <t1, t2> in m):
    if (t1 = t2):
      use t1 as new type
    else if (subtype(t1, t2)):
      use t2
    else if (subtype(t2, t1)):
      use t1
    else if (∃ t3 such that subtype(t1, t3) && subtype(t2, t3)):
      use t3
    else:
      if (generalize(g', t1, t2) returns a unique construction):
        use generalize(g', t1, t2)
      else:
        use createNewCategory(t1, t2)

  if (sameLength(c1, c2)):
    return one new concrete construction
  else:
    return two new concrete constructions
```

Figure 4.15 The procedure for generalizing two constructions c1 and c2. If their constituents do not have an existing parent type, a recursive generalization is triggered.

Lifting form and meaning constraints

After constituents are aligned and the new constituent types are determined, the form constraint, meaning pole type, and meaning constraints of the general construction have to be determined. As the new concrete construction is intended to eventually replace the specific constructions, all constraints are expected to be lifted to the new construction.

The least specific set of form constraints from the two candidates are retained. The meaning pole type, as expected, is the most specific common ancestor of the two original meaning pole, and the same goes for each contextual constraint. Slot chains are re-written using new constituent names, new local names for lifted evoked elements, and mapped role names if the meanings of constituents have been generalized. An seen in the resulting construction CAT001-CAT002 construction in the WO3-CHI1 (I eat) / NI3-HE1 (you drink) example, the new meaning pole type is INGESTION, the contextual constraints on its `rd1` is relaxed and the speech_act constraint is dropped altogether.

In rare instances, certain constraints can no longer be representable given the new constituent types since the required roles are no longer accessible given a more general constituent meaning. In these situations, the generalization operation is aborted.

Generalizing across two constructions with differing number of constituents

In order to compensate for sparse input data and frequent argument omission, the model implements an aggressive generalization algorithm that also works when two constructions differ in length by one constituent, such as WO3-CHI1 (I eat) and XI1XI1-CHI1-YAO4 (XiXi eat medicine). Much of the algorithm remains unchanged, and the structural alignment process proceeds with only the shared constituents. A generalized version of each of the constructions is created, leaving most of the constraints intact but replacing some constituents with the generalizations.

In this example, one new category CAT003:{WO3-N, XI1XI1-N} is created and two additional new concrete constructions CAT003-CHI1 and CAT003-CHI1-YAO4 are learned. These two concrete constructions resemble their corresponding specific constructions as shown in Figure 4.16.

```
Construction CAT003-CHI1
   subcase of CLAUSE
   constructional
      constituents
         c0 : CAT004
         c1 : CHI1-V
   form
      constraints
         c0.f meets c1.f
   meaning : EAT
      evokes RD as rd0
      evokes RD as rd1
      evokes DISCOURSE_SEGMENT as DS
      constraints
         self.m <--> c1.m
         c0.m <--> c1.m.ingester
         rd0 <--> c0.rd
         rd0.referent <--> c1.m.ingester
         rd1.referent <--> c1.m.ingested
         rd1.ontological_category <-- @Rice
         rd1.discourse_role <--
            @Attentional_Focus
         DS.speech_act <-- EXPLAINING
```

```
Construction CAT003-CHI1-YAO4
   subcase of CLAUSE
   constructional
      constituents
         c0 : CAT004
         c1 : CHI1-V
         y2 : YAO4
   form
      constraints
         c0.f meets c1.f
         c1.f meets y2.f
   meaning: EAT
      evokes RD as rd0
      evokes RD as rd1
      evokes DISCOURSE_SEGMENT as DS
      constraints
         self.m <--> c1.m
         c0.m <--> c1.m.ingester
         y2.m <--> c1.m.ingested
         rd0 <--> c0.rd
         rd0.referent <--> c1.m.ingester
         rd1.referent <--> c1.m.ingested
         rd1.ontological_category <-- @Medicine
         rd1.discourse_role <--
            @Attentional_Focus
         DS.speech_act <--
            REQUESTING_ACTION
```

Figure 4.16 Creating two new concrete constructions when constructions of differing lengths are generalized. In this case, generalization across WO3-CHI1 **(I eat) and** XI1XI1-CHI1-YAO4 **(XiXi eat medicine) lead to a new category with** WO3-N **and** XI1XI1-N **as members and two new constructions** CAT003-CHI1 **and** CAT003-CHI1-YAO4**.**

Competition between general and specific constructions

In this learning model, there are 3 options for the more (lexically-) specific constructions such as WO3-CHI1 (I eat) and NI3-HE1 (you drink) after they have been generalized: (1) they can be left unchanged, essentially letting lexically-specific constructions compete with the newly learned CAT001-CAT002 ({I, you} {eat, drink}) construction, (2) they can be made subcases of the new construction, or (3) they can be removed from the grammar at the time of generalization. There are competing claims in the literature with respect to whether the lexical constructions continue to exist in the grammar, though there is strong evidence that they do (Bybee & Scheibman., 1999). The initial implementation uses option (1), but the other options will be left as experiments to be tested on the model.

Summary: the generalization operation

The current formulation of the generalization process separates the issue of retrieval strategy (selecting constructions from the grammar to try to generalize — shared constituents or shared semantics) from that of analogy (deciding whether a pair of constructions is similar enough to generalize). This learning model provides a formalized framework that allows both to be manipulated in computational experiments, and in particular to examine constructional generalization from six perspectives:

1. form: ordering constraints are used as a filter on whether two constructions are candidates for generalization.
2. meaning: while the semantics of a construction acts as a filter for retrieving possible generalization candidates, detailed structural alignment of semantic constraints is also necessary for two constructions to generalize
3. constituent structure: the model allows for conservative generalization where constituent structures have to match perfectly as well as aggressive generalization where partial matches are permitted.

4. grammatical context: triggered generalization, including category formation, necessarily take into account the syntactic context of a constituent because these generalizations are brought about by some other constructions that use it
5. situational context: a fixed number of contextual constraints are allowed to be relaxed.
6. discourse context: a fixed number of discourse constraints are allowed to be relaxed.

Chapter 5.

Refining Previously Learned Constructions

The last chapter gave the basic mechanisms with which grammatical constructions are learned, but in a noise-prone learning environment learning mistakes are unavoidable. Incorrect context fitting may lead to incorrect meaning relations in constructions being posited, and inattention to unstressed grammatical particles may lead to important syntactic distinctions being lost. Spurious constructions with overly specific contextual constraints could be learned, while at other times generalizations may be too conservative to be useful. The model therefore provides five means of refinement and correction: constituent omission, construction revision, category merge, category expansion, and grammar decay.

5.1 Detecting the need for refinement

As was briefly mentioned in Chapter 4, misalignment during the structural comparison process indicates a need for refinement for one or both of the constructions. Consider the following pair of learned constructions, YUE4LIANG4-CAT006 (Moon - CAT006) and NI3-CAT006 (you - CAT006), where CAT006 are made up of verbs such as SI1-V (tear), BAI1-V (rip apart), HUI3-V (damage). They both refer to the same scene: it turned out that in the scene, the child came across a page in a picture book which she had accidentally ripped some time ago — the word *yue4+liang4* (moon) was used metonymically for the page on which the moon was drawn. By aligning the constituents in the two constituents (y0 with n0, and c1 with c1) the learner realizes that the two constructions look syntactically similar (a verb preceded by a noun) while the meaning relations are in conflict. In YUE4LIANG4-CAT006, the moon is the undergoer of the TEAR action, whereas in NI3-CAT006, the addressee is the actor in the TEAR action. The rest of the situational contexts look largely compatible so there is reason to suspect that there are additional grammatical cues that the learner has missed which will help differentiate the two meanings. This pair of constructions is put on a revision watch list.

```
Construction YUE4LIANG4-CAT006
   subcase of CLAUSE
   constructional
      constituents
         y0 : YUE4LIANG4-N
         c1 : CAT006
   form
      constraints
         y0.f meets c1.f
   meaning : TEAR
      evokes DISCOURSE_SEGMENT as DS
      evokes RD as rd1
      evokes RD as rd0
      constraints
         c1.m.undergoer <--> y0.m
         rd0.referent <--> c1.m.undergoer
         rd1.referent <--> c1.m.actor
         rd0.ontological_category <-- @Moon
         rd1.ontological_category <-- @Child
         rd1.discourse_role <-- @Addressee
         DS.speech_act <-- EXPLAINING
```

```
Construction NI3-CAT006
   subcase of CLAUSE
   constructional
      constituents
         n0 : NI3_VARIANT
         c1 : CAT006
   form
      constraints
         n0.f before c1.f
   meaning : TEAR
      evokes DISCOURSE_SEGMENT as DS
      evokes RD as rd1
      evokes RD as rd0
      constraints
         c1.m.actor <--> n0.m
         rd0.referent <--> c1.m.undergoer
         rd1.referent <--> c1.m.actor
         rd1 <--> n0.rd
         rd0.ontological_category <-- @Moon
         DS.speech_act <-- EXPLAINING
```

Figure 5.1 Two constructions, YUE4LIANG4-CAT006 **(moon {tear, rip, damage}) and** NI3-CAT006 **(you {tear, rip, damage}), that align in form but not meaning. The pair is a candidate for revision.**

The same kind of structural alignment takes place between constructions of unequal length as well, such as the pair in Figure 5.2. The first , GEI3-CLAUSE, contains just the verb GEI3_VARIANT-V (give) whereas the second, GEI3-WO3, contains both the verb and a noun WO3 (I) who is the recipient. Aligning the verb with the other verb in these two constructions, the learner found that the two constructions have the same meaning (a GIVE scene) as well as contextual constraints (a Child as the giver and a Pen as the theme) except for the difference of one constituent (w1). It is possible that w1 is either optional or omissible, but the learner cannot make a conclusion on the basis of the two constructions themselves — even if GEI3-CLAUSE is in the grammar alongside GEI3-WO3, GEI3-CLAUSE can be used as a constituent of another construction which supplies the additional arguments. The learner thus needs to wait and see how the shorter construction is used. The pair of construction is put on an omission watch list.

```
Construction GEI3-CLAUSE
    subcase of CLAUSE
    constructional
        constituents
            g0 : GEI3_VARIANT-V
    meaning : GIVE
        evokes DISCOURSE_SEGMENT as DS
        evokes RD as rd1
        evokes RD as rd0
        evokes RD as rd2
        constraints
            self.m <--> g0.m
            rd0.referent <--> g0.m.giver
            rd1.referent <--> g0.m.theme
            rd2.referent <--> g0.m.recipient
            rd0.ontological_category <-- @Child
            rd0.discourse_role <-- @Addressee
            rd1.ontological_category <-- @Pen
```

```
Construction GEI3-WO3
    subcase of CLAUSE
    constructional
        constituents
            g0 : GEI3_VARIANT-V
            w1 : WO3
    form
        constraints
            g0.f meets w1.f
    meaning : GIVE
        evokes RD as rd0
        evokes RD as rd1
        evokes RD as rd2
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> g0.m
            rd0.referent <--> g0.m.giver
            rd1.referent <--> g0.m.theme
            rd2.referent <--> g0.m.recipient
            w1.m <--> rd2.referent
            w1.rd <--> rd2
            rd0.ontological_category <-- @Child
            rd0.discourse_role <-- @Addressee
            rd1.ontological_category <-- @Pen
            rd2.ontological_category <-- @Mother
            rd2.discourse_role <-- @Speaker
            DS.speech_act <-- EXPLAINING
```

Figure 5.2 Two constructions that differ in length by one constituent. The different constituent, w1, may be omissible but the learner needs to wait till GEI3-CLAUSE is used again — alone — to be sure.

5.2 Construction revision

When a construction on the shortlist for revision is used in the analysis of an utterance, the revision operation is triggered. Its primary function is to look for additional grammatical cues that might differentiate two constructions that look the same syntactically and mean different things. These cues may be in the form of additional function words or content words. We turn again to the first example given in the last section to illustrate the operation, but real data is always much more complicated. The following is the dialogue from which the constructions are learned:

1	*yue4+liang4*	*si1*	*er(le)*			(the moon is torn)
	moon	tear	PERF			
2	*zhei4*	*si1*	*le*			(this is torn)
	this	tear	CRS/PERF			
3	*ni3*	*kan4*				(you see)
	you	see				
4	*ni2*	*gei3*	*si1*	*le*		(you tore (it))
	you	CV	tear	CRS/PERF		
5	*ba3*	*yue4+liang4*	*gei3*	*si1*	*le*	(you) tore the moon
	CV	moon	GEI3	tear	CRS/PERF	

Recall that the need for revision arose because the two constructions create an ambiguity in the grammar: the noun-verb pattern seems to allow both the actor and the undergoer readings. It turns out both readings are indeed confirmed by data but the difference is subtle. As can be seen from (1) and (2), the undergoer can be the subject of a sentence but the construction actually denotes undergoer-state. This sentence-final particle *le*, which denotes both a current relevant state (CRS) and a perfective aspect, is important in turning the verb into a stative description of the undergoer and making the sentence grammatical.

A different way of placing the undergoer preverbally without the stative reading is by fronting it using the *ba3* coverb as in (5).[29] The learner needs to pay attention to the pre-nominal *ba3* particle which clearly marks the moon as the object. For the purpose of improving the grammar, it suffices for the learner to incorporate either the aspect particle *le* or the object-marking coverb *ba3* into the construction. With no knowledge about specific function words,

[29] In some local dialects of Mandarin an additional *gei3* particle is also used to emphasize affectedness. It is not exactly a passive but is most often used when the object appears preverbally.

however, how is the learner to tell these two useful particles apart from the other unknown words in the sentence?

The model uses a combination of adjacency constraints and usage statistics to select new constituents to serve as grammatical cues in differentiating two conflicting constructions. When a construction on the revision watch list is used in an analysis, the learner looks for other constructions in the analysis that are immediately next to or fall inside the span of the watched construction — the latter happens because of the analyzer's ability to do robust parsing and skip over unconnected roots. A usage-based filter is then applied to all those adjacent constructions. There are a number of possible frequency-based heuristics, the choice among which is an empirical or an engineering issue. For model simplicity, we put a threshold on the constructional bigram. For each adjacent construction that exceeds the threshold, a new version of the construction in question with the extra constituent is created.

```
given Revisable construction r, Analysis a

Set<Construction> C = constructions in a adjacent to r
Set<Construction> O = old constituents of r

for c in C:
  for o in O:
    bigram(c,o) = P(c|o) if o is before c in a, P(o|c) o.w.
    if (bigram(c, o) > threshold):
      create one new construction with constituents {O, c}
```

Figure 5.3 Using adjacency and frequency information to sift out grammatical cues in revision.

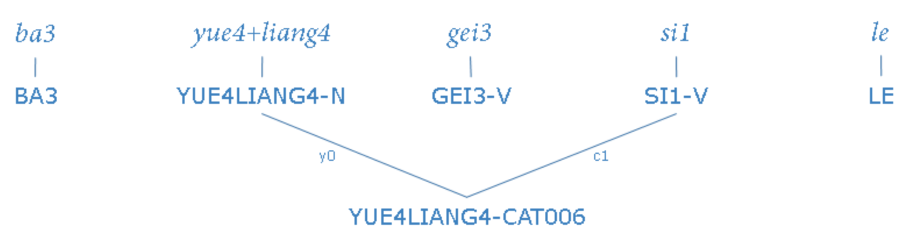

Figure 5.4 The expected analysis of (5) using YUE4LIANG4-CAT006. BA3, GEI3-V, and LE are all adjacent to YUE4LIANG4-CAT006 or its constituents.

Carrying this operation out on YUE4LIANG4-CAT006 the next time it is used to analyze (5), one constructional analyses will look like the above.[30] Since all of BA3, GEI3-V and LE are potential new constituents to add to the YUE4LIANG4-CAT006 construction due to adjacency, the bigrams P(YUE4LIANG4-N | BA3), P(GEI3-V|YUE4LIANG4-N), P(SI1-V|GEI3-V), and P(LE|SI1-V) are examined to determine the appropriateness of each as a new constituent. If BA3 passes the criterion, a new construction BA3-YUE4LIANG4-CAT006, as shown below, will be created.

```
Construction BA3-YUE4LIANG4-CAT006
    subcase of CLAUSE
    constructional
        constituents
            b0 : BA3
            y1 : YUE4LIANG4-N
            c2 : CAT006
    form
        constraints
            y0.f meets c1.f
    meaning : TEAR
        evokes DISCOURSE_SEGMENT as DS
        evokes RD as rd1
        evokes RD as rd0
        constraints
            c1.m.undergoer <--> y0.m
            rd0.referent <--> c1.m.undergoer
            rd1.referent <--> c1.m.actor
            rd0.ontological_category <-- @Moon
            rd1.ontological_category <-- @Child
            rd1.discourse_role <-- @Addressee
            DS.speech_act <-- EXPLAINING
```

Figure 5.5 A revised version of YUE4LIANG4-CAT006 with an additional particle BA3 is created and added to the grammar.

5.3 Constituent omission

The omission watch list, as the reader may recall, contains pairs of constructions that mean the same thing (in context) but differ in their number of constituents. When the learner encounters an analysis in which the shorter construction in a pair is used, the omission operation

[30] The *gei3* is likely to be misinterpreted to be a verb, depending on the starting lexicon. How the function word sense of *gei3* can be separated from the most frequently used verb sense leaves another story to be discussed, though the short answer is a combination of collocation frequency and semantic-context fit.

is triggered. The usage of the shorter construction is by itself insufficient evident for an omissible or optional constituent — the construction maybe be embedded in a larger one that supplies the "missing" meaning — so semantic verification must be carried out.

Specifically, semantic roles that are connected with the differing constituent in the longer construction must be verified in the current analysis that uses the shorter construction. If they are not filled in constructionally, indicating that the meaning is either only contextually specified or has to be inferred, it implies that the differing constituent does not need to be overtly mentioned. If the corresponding semantic role(s) are core roles, the constituent is notated as omissible, else the constituent is marked optional using the optional keyword. In essence, the optionality or omissibility of the constituents of a construction is determined by contrasting it with the usage of a minimally different construction.

```
given Grammar g, Pair<Construction> {shorter, longer},
differing constituent c, Analysis a

find semantic roles R connected to c in longer
look up R in a

if (none of R is filled in constructionally):
  if (one or more of R are core roles):
    c is marked omissible in longer
  else:
    c is marked optional in longer

remove shorter from g
```

Figure 5.6 When the shorter construction of a marked pair is used in an analysis, an effort is made to ensure that no other constructions are filling in the corresponding semantic roles before making the differing constituents optional or omissible.

Taking as a concrete example the pair GEI3-CLAUSE and GEI3-WO3 introduced in Figure 5.2, the learner's job is to determine if the constituent w1 in the latter is an optional or omissible. These three unification constraints,

rd2.referent <--> g0.m.recipient

w1.m <--> rd2.referent

w1.rd <--> rd2

tells the learner that w1 is connected to the recipient role of the GIVE schema. If GEI3-CLAUSE is used again to analyze another utterance, the recipient role is examined in the resulting analysis. If it is filled in constructionally, that is, if there is a path from another construction in the analysis to the recipient slot, nothing can be concluded from the learning input. However, if it is not constructionally filled, the learner deduces that w1 need not be overtly expressed. Since recipient is a core role, the learner marks w1 as omissible.

5.4 Category merge

Every so often, a new grammatical category is created over other existing grammatical categories due to generalization operations. Take as an example CAT016 below, consisting of the verbs CUO1-V (Mash) and RENG1-V (Throw), and this category is used by, say, construction CXNA. Another construction, say, CXNB, that uses the verb TI1-V (Kick) comes along and is generalized with CXNA, creating a new CAT033 over the existing CAT016. One obvious optimization is to merge CAT013 into CAT033.

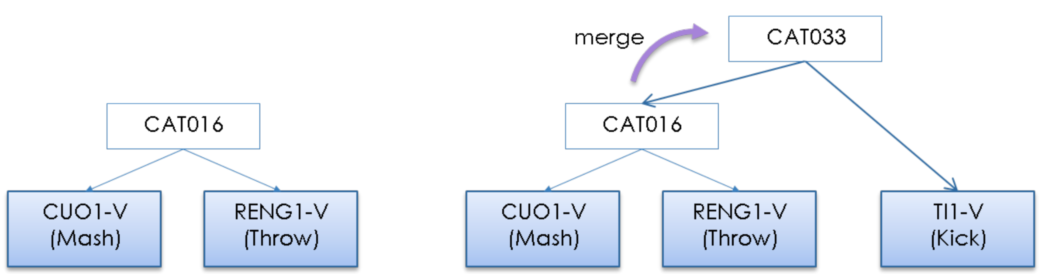

Figure 5.7 A generalization operation may cause a new category to be created over an existing one. A natural optimization is to merge the two categories.

A more general scenario is when the grammar may be better served by combining two or more categories that already share some members, as illustrated in Figure 5.8. Category merge is a lot more like generalization than it seems at first because abstract constructions may have constituents and semantic constraints as well as evoked roles. When two categories are merged the shared structures have to be extracted. Unfortunately, because a category also acts like an interface to constructions which use it, guaranteeing the semantic contents of its members, merging two interfaces that make different guarantees (or have different names for the same semantic roles) will break the 'contracts' with their users.

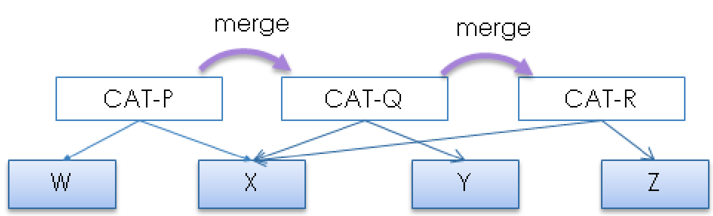

Figure 5.8 A more general problem of merging multiple categories. Because abstract constructions contain constituency, form, and meaning structures, when performing a merge between categories, the shared structures have to be retained and unshared ones pushed back downwards to the respective category members. This may potentially break a category's "contract" or guarantees with users of the category, but a general solution is beyond the scope of the current work.

There is no general solution to automatically fix the grammar if the semantic guarantees of a linguistic category are altered, and the degree of optimization is mostly of engineering consequence. As such, category merge is only partially supported by the learner in cases where the least restrictive of the merged category suffices as an interface for all the resulting members. In other words, category merge between a set of grammatical categories is performed pairwise with no attempt at reconciliation between role names until all categories are merged or until the operation results in broken 'contracts' between categories and their users.

Pairwise category merge is carried out using the same algorithm that creates a linguistic category in the first place. By generalizing over the two categories, only the shared components are retained in the new category (the *placeholder* category in the following pseudo-code), which is to be used in the grammar in place of the original ones. The placeholder category, however, may have fewer constraints than in either of the original categories. Since multiple inheritance is allowed in ECG, some of these constraints may still be inherited by members of the original categories from other parents. Those that are not, however, are pushed back downwards.

```
given Grammar g, Construction cat1, Construction cat2

g' = g.makeCopy()
placeholder = generalize(g, cat1, cat2)

Set<Constraints> T1 = all constraints in cat1 not in placeholder
for construction member in cat1:
  for each constraint t in T1:
    if (t is not in member through other means of inheritance):
      add t back to member

Set<Constraints> T2 = all constraints in cat2 not in placeholder
for construction member in cat2:
  for each constraint t in T2:
    if (t is not in member through other means of inheritance):
      add t back to member

remove cat1, cat2 from g'
rename placeholder as cat1 and add to g'
if (well-formed(g')):
  accept g'
```

Figure 5.9 The category merge operation makes use of the generalization algorithm that creates a linguistic category in order to extract the shared structures between two categories. The unshared structures are examined and pushed back down to the members if necessary.

5.5 Category expansion

The generalization operation described in Section 4.2 is a very conservative generalization operation. The generalization between two constructions creates very narrow linguistic categories that may consist of only two items, and until the categories get broader, the use of the generalization is very limited. This learning model therefore allows for a specific kind of semantics-based category expanion that is triggered when the category is used or otherwise touched on by another learning operation. Continuing with the example introduced in Figure 5.7, we have CAT033 with three members now after the merge: CUO1-V (Mash), RENG1-V (Throw), and TI1-V (Kick). The learner looks for ways to further extend the category by checking the schema lattice. It attempts to extend CAT033 to constructions with meanings that are sibling to MASH, THROW, and KICK, namely, DA3-V (Beat), ZHUI4-V (Pound), PAI1-V (Slap), and the like.

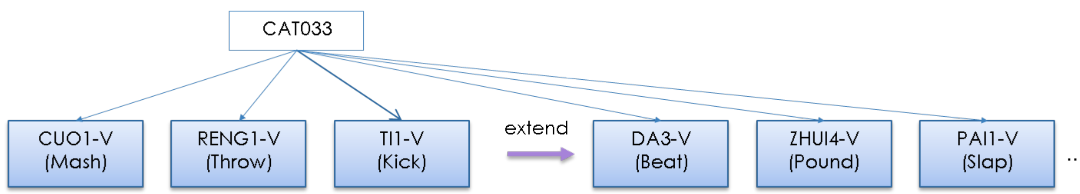

Figure 5.10 CAT033 **from the last example is now being semantically extended to include other verbs of agentive impact.**

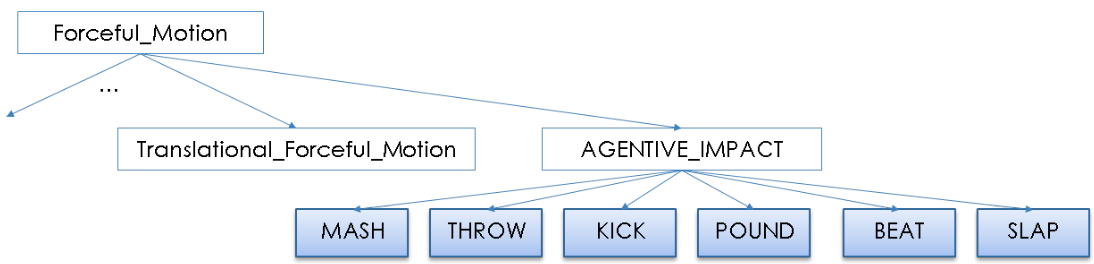

Figure 5.11 Portion of the semantic lattice showing semantic siblings of MASH, THROW, **and** KICK

Given this intuition, the learner has to proceed with caution in making sure that constructions recruited to be new members conform to the additional semantic constraints which the category poses. That is, each new member must already be subsumed both syntactically and semantically by the category (see Section 3.4 for how the learner determines subsumption between constructions). The whole algorithm goes as follows:

```
given Grammar g, Construction cat

Schema m = meaning pole type of cat
Set<Schema> S = siblings(m)
Set<Construction> C = constructions whose meaning is among S

for Construction c in C:
  if (subsumes(cat, c)):
    add cat to parents of c
    Set<Constraints> T = constraints in c already in cat
    remove T from c
```

Figure 5.12 The category expansion operation is a semantically driven way to increase the generalization capacity of the grammar.

5.6 Grammar decay

At each learning step, old, unused concrete constructions are purged. Each construction is timestamped with a date of last modification which is initially set when the construction is created and is updated each subsequent time the construction is updated (e.g. modification of constructional parent). The current 'date' increments each time any modification is made to the grammar, effectively numbering the different versions of grammar that the learning model has gone through. Constructions that are learned long time ago (older than a user-set duration) and are not entrenched (usage frequency less than a user-set number) are removed from the grammar. Any resulting orphaned categories (categories that are no longer used by any constructions) are removed as well.

Chapter 6.

Keeping Statistics on a Grammar

If constructions provide the structures that guide the understanding of language, then statistical information on these constructions expresses preferences on these structures. Bryant, Bybee, and Gahl have all separately argued that these sorts of probabilistic knowledge should be considered a part of grammatical knowledge (Bryant, 2008b; Bybee, 2006; Gahl & Garnsey, 2004; 2006). That is the approach adopted in this work.

The importance of statistical information to this model is illustrated by how the use (and non-use) of newly learned constructions greatly affects the trajectory of learning. Usage matters in more ways than one: the use of a construction may trigger generalization (see Section 4.2), while extended non-use leads to its removal from the grammar (see Section 5.6). A third, arguably more impactful, way in which usage matters is through competition during the analysis process between specific and general constructions. Entrenchment, or frequent usage, of a specific construction increases its likelihood of being used again in the future. Comparatively, its generalization becomes less likely to be used, to the point where the generalization may eventually be lost to grammar decay. This is a trade-off that a grammar learning model must cope with as long as it allows specific constructions to co-exist alongside their generalizations. We will return again to this idea in our final discussion in Chapter 9.

Consider the generalization between WO3-CHI1 (I eat) and NI3-HE1 (you drink) in Section 4.2 that leads to the CAT001-CAT002 construction and two narrowly defined categories CAT001 and CAT002, each consisting of two members. In the simplest implementation, the two specific constructions are retained in the grammar along with the new general construction. Notice, however, that until either category is expanded, the CAT001-CAT002 construction is only slightly more general than the two more specific constructions combined. The learner only gains the ability to analyze *wo3 he1* (I drink) and *ni3 chi1* (you eat) in addition to *wo3 chi1* (I eat) and *ni3 he1* (you drink). If the learning input is sufficiently large and diverse, we may expect the generalization to eventually dominate in usage. However, in practice, even though standard smoothing procedures are applied to the usage counts to assign non-zero probabilities to new constructions, the input data to the learning model is so sparse that new constructions can benefit from some amount of additional probability mass to help them gain traction in usage. At the same time, omission probabilities must also be properly updated in order for the analyzer to make the right trade-off between a shorter construction with fewer constituents and a longer, semantically richer construction with omitted constituents. Getting the learned constructions to be used correctly in future language understanding is a big part of the learning process.

As explained in Section 2.3, the analyzer takes as input these probabilities for all constructions α, constituents β of α, and frames f:

- The locality probabilities, $\mathrm{P}(expressed_\beta \mid \alpha)$ and $\mathrm{P}(local_\beta \mid expressed_\beta, \alpha)$, which represent the probability that constituent β of construction α is expressed and the probability that constituent β of α is expressed locally given that it is expressed. An example is the probability that constituent c0 in the CAT001-CAT002 construction is not expressed (i.e. omitted).
- The constructional filler probability, $\mathrm{P}(filler_\beta \mid expressed_\beta, \alpha)$, which represents the probability that constituent β of construction α is filled by construction $filler_\beta$. An

example is the probability that constituent c0 in the CAT001-CAT002 construction is filled by WO3-N.

- The semantic presence probabilities, $\mathrm{P}(\mathit{filled}_{role} \mid \mathit{role}_f, f)$, which represents the probability that a particular role role_f of frame f is filled. An example is the probability that the ingested role of the INGESTION schema being filled.
- The semantic role probabilities, $\mathrm{P}(\mathit{role}_f \mid \mathit{filler}, f)$, which represents the probability that a particular filler *filler* is assigned to the role role_f given frame *f*. An example is the probability that a @Peach fills the ingested role of the INGESTION schema.

The two semantic probabilities are taken to be prior knowledge and remain stable throughout learning to simplify the model. The two constructional probabilities are taken to be learned linguistic knowledge and are accumulated through exposure to linguistic input. This chapter describes how they are updated throughout learning. We will focus first on how the locality and constructional filler counts are tracked throughout usage, then discuss how each learning operation explicitly manipulates the counts to encourage the use of new constructions. The method for smoothing the occurrence counts for the analyzer's use will be discussed in the last part of this chapter.

6.1 Updating statistics through usage

Given a static, stable grammar, statistics on existing constructions are gathered in a fairly straightforward manner as the model encounters learning input. Both lexical and constructional unigram and bigram probabilities for words w_i and constructions α_i are gathered: $\mathrm{P}(w_i)$, $\mathrm{P}(\alpha_i)$, $\mathrm{P}(w_i \mid w_{i-1})$, $\mathrm{P}(\alpha_i \mid \alpha_{i-1})$. The corresponding unigram and bigram frequencies, $\mathrm{freq}(w_i)$,

$\mathrm{freq}(\alpha_i)$, $\mathrm{freq}(w_{i-1}, w_i)$, $\mathrm{freq}(\alpha_{i-1}, \alpha_i)$ are updated based on the constructional tree of the best analysis for each utterance. In addition, the locality counts and constructional filler counts from which the locality probabilities and constructional filler probabilities are derived are updated throughout learning. For each construction α in the best analysis, the learner keeps track of:

- $\mathrm{C}(locality_\beta, \alpha, \beta_j^\alpha)$: the number of times its j^{th} constituent β_j^α has locality $locality_\beta$, where $locality_\beta$ can take on one of four constant values:
    - *LOCAL:* β_j^α is expressed locally
    - *EXTRAPOSED:* β_j^α is expressed nonlocally
    - *OMITTED :* a non-optional constituents β_j^α is omitted
    - *UNFILLED*: an optional constituents β_j^α is unfilled
- $\mathrm{C}(filler_\beta, \alpha, \beta_j^\alpha)$: for each constituent β_j^α that is filled (either locally or extraposed), the number of times constituents β_j^α is filled by $filler_\beta$

Notice that only concrete constructions can be fillers of constituents. Additionally, a subtle choice is made with this update process with regard to construction generalizations, namely, that the counts of the more lexically-specific constructions are updated independently of those of the subsequently derived generalizations, pitting the constructions directly against one another during use. (An alternative procedure may be to assign partial credit to generalization parents and/or children as well). The implications of this choice will be discussed in Chapter 9.

6.2 Updating statistics through learning

Given that constructional statistics express preferences on grammatical structures, we expect these preferences to shift over the course of learning, possibly due to explicit interventions from the learning operations. With little past empirical work to inform these choices, this section is an attempt to systematically outline how statistical information is updated per learning operation.

For any given construction α, let us define a userυ as a construction which has one or more constituents that can expand directly into α (in other words, a construction with one or more constituents whose type constraint is an ancestor of α). For each new construction α *with constituents* $\beta_1^\alpha \dots \beta_n^\alpha$ and *potential users* $\upsilon_1 \dots, \upsilon_m$, these distributions available for adjustment during learning:

- $\mathrm{P}(\alpha)$: the constructional unigram probability of α
- $\mathrm{P}(\alpha|\varphi)$ and $\mathrm{P}(\varphi|\alpha)$: the constructional bigram probabilities for all constructions φ in the grammar, i.e. the probability of construction α given that the preceding construction is φ, and the probability of construction φ given that the preceding construction is α.
- $\mathrm{P}(filler_\beta|\alpha, \beta_j^\alpha)$: the probability that the j^{th} constituent β_j^α of α is filled by $filler_\beta$. For brevity we refer to this as the **internal** constructional filler distributions
- $\mathrm{P}(\alpha|\,\upsilon, \beta_k^\upsilon)$: the probability that α fills the k^{th} constituent β_k^υ of a user υ. For brevity we refer to this as the **external** constructional filler distributions.
- $\mathrm{P}\big(expressed_\beta|\alpha\big)$ and $\mathrm{P}\big(local_\beta|expressed_\beta, \alpha\big)$: the locality distributions for each constituent β_j^α

From a language understanding accuracy standpoint, the best performance can be obtained by re-estimating all these parameters from the entire set of learning data every time a new construction is added. This is obviously neither efficient nor cognitively plausible. An alternative is for the model to update in a principled way the occurrence counts for the new construction based on existing ones, for each learning operation. The basic update calculations, which will be explained in detail over the next few subsections, are derived from the assuming each learned construction is correct and supported by all previously encountered data, such that a new composition immediately assumes a count of 1 (that is, the current utterance which triggered the composition counts as one piece of evidence for the new composition being used) and a new generalization assumes the total counts of the specific constructions (that is, all previous utterances that are covered by the specific constructions are now evidence for the new generalization). Obviously, the input data is noisy and the learner can make mistakes, so a parameter $0 \leq \gamma \leq 1$ is used as a discount factor for the count updates to reflect the uncertainty in the learning process.

The idea of constituent mapping is important throughout the explanation of the update mechanism. Recall from Section 3.4 that a **constituent mapping** is a one-to-one correspondence (null permitted) between constituents of α and constituents of β. Constituent maps exist (1) between a general construction and the specific constructions from which it is created, (2) between a revision and the construction that is revised, and (3) between a construction with omissible or optional constituents and the pair of constructions from which the omission / optionality status is deduced. Examples of these will be given over the next sections.

Composition

Composed constructions are made up of smaller existing constructions. Assuming that the composition is correct, the utterance which leads to the composition is evidence that this composition is used. Therefore, the undiscounted constructional unigram frequency for each new composed construction, $\mathrm{freq}(\alpha)$, is 1. There is little basis in the kept statistics for estimating the constructional bigrams for a new composition, so they remains unchanged.

The constituents β of the new composition α are narrowly type restricted to the constructions present in the constructional analysis at the time of the composition operation, and the internal constructional filler occurrences $\mathrm{C}(filler_{\beta},\alpha,\beta_j^{\alpha})$ are incremented by γ each. The new construction α, being a subcase of either PHRASE or CLAUSE, may be used as a constituent in another construction ν that has a constituent of those types. The external constructional filler occurrences $\mathrm{C}(\alpha,\nu,\beta_k^{\nu})$ for each potential user ν may therefore also be adjusted. Whether to update this count is a design decision of the learning model. Incrementing this count is equivalent to saying that the newly learned construction α has been observed in the constructional context of another construction ν. It is not currently implemented since such an update does not seem to be warranted by the data. Finally, since each constituent β of the new composition α comes from a construction used in the analysis, its *LOCAL* count is incremented by γ.

These updates are illustrated in Figure 6.1 below using the composed construction XI1XI1-CHI1-YAO4 (Xixi eat medicine) from Figure 4.6, which has constituents x0, c1 and y2. We take the utterance from which it was learned, *xi1+xi1 chi1 yao4*, as evidence of its use.

COMPOSITION
new construction: XI1XI1-CHI1-YAO4 (Xixi eat medicine)
composed from: XI1XI1-N, CHI1-V, YAO4-N

constructional unigram frequency
$freq$(XI1XI1-CHI1-YAO4) = $\gamma \cdot 1$

constructional bigram frequency
no change

internal constructional filler counts
C (XI1XI1-N, XI1XI1-CHI1-YAO4, x0) = $\gamma \cdot 1$
C (CHI1-V, XI1XI1-CHI1-YAO4, c1) = $\gamma \cdot 1$
C (YAO4-N, XI1XI1-CHI1-YAO4, y2) = $\gamma \cdot 1$

external constructional filler counts
no change

locality counts
C ($LOCAL$, XI1XI1-CHI1-YAO4, x0) = $\gamma \cdot 1$
C ($LOCAL$, XI1XI1-CHI1-YAO4, c1) = $\gamma \cdot 1$
C ($LOCAL$, XI1XI1-CHI1-YAO4, y2) = $\gamma \cdot 1$.

Figure 6.1 Example of updating the statistics after a composition operation.

Generalization

The updates after a constructional generalization are a bit more complex. Taking first the simple case when the specific constructions are all equal length and one new concrete, general construction α_{new} from a set of more specific constructions $\alpha_1 \ldots \alpha_i$ is created. Since the generalization α_{new} subsumes $\alpha_1 \ldots \alpha_i$ and is intended to eventually replace them in the grammar, on most statistics it gets the sum of the counts held by the specific constructions, treating the evidence for the specific constructions as evidence for itself.

More precisely, the constructional unigram frequency of the general construction, freq(α_{new}), is the discounted sum of the unigram frequencies of the specific constructions,

$\gamma \sum_i \mathrm{freq}(\alpha_i)$. The constructional bigram frequency of the new construction $\mathrm{freq}(\varphi, \alpha_{new})$ is the discounted sum of the bigram frequencies of the specific constructions, $\gamma \sum_i \mathrm{freq}(\varphi, \alpha_i)$, for all constructions φ. The same applies for the other bigram frequency $\mathrm{freq}(\alpha, \varphi)$.

The internal constructional filler count for constituent β_j^α in α_{new}, $\mathrm{C}(filler_\beta, \alpha_{new}, \beta_j^\alpha)$, is given by the discounted sums of the internal constructional filler counts of the corresponding constituents in the specific constructions, $\gamma \sum_i \mathrm{C}(filler_\beta, \alpha_i, \beta_j^{\alpha_i})$. The intuition is that whatever can act as a constituent of the specific construction can act as a constituent of the general construction, too. The external constructional filler count of construction α_{new} in constituent β_k^υ in user υ, $\mathrm{C}(\alpha_{new}, \upsilon, \beta_k^\upsilon)$, is given by the discounted sums of the external constructional filler counts of the specific constructions, $\gamma \sum_i \mathrm{C}(\alpha_i, \upsilon, \beta_k^\upsilon)$. The intuition here is that wherever the specific constructions are used, the general construction can be used, too. This is feasible because both the specific constructions and the generalization(s) share the same constructional parents (PHRASE or CLAUSE). Finally, the locality count of the general construction, $\mathrm{C}(locality_\beta, \alpha_{new})$, is obtained likewise by summing across the locality counts of the specific constructions, $\gamma \sum_i \mathrm{C}(locality_\beta, \alpha_i)$.

These update operations are illustrated in Figure 6.2 using the CAT001-CAT002 construction from Figure 4.12 which is generalized from WO3-CHI1 and NI3-HE1. The two categories created concurrently are CAT001:{WO3, NI3} and CAT002:{CHI1, HE1}.

CONCRETE GENERALIZATION

new construction: CAT001-CAT002

generalized from: WO3-CHI1 (I eat) and NI3-HE1 (you drink)

created categories: CAT001:{WO3, NI3}, CAT002:{CHI1, HE1}

final constituent mapping:

<CAT001-CAT002, c0>: <WO3-CHI1, w0>, <NI3-HE1, n0>

<CAT001-CAT002, c0>: <WO3-CHI1, w0>, <NI3-HE1, n0>

constructional unigram frequency

freq(CAT001-CAT002) = $\gamma \cdot$ (*freq*(WO3-CHI1) + *freq*(NI3-HE1))

constructional bigram frequency

freq(φ, CAT001-CAT002) = $\gamma \cdot$ (*freq*(φ, WO3-CHI1) + *freq*(φ, NI3-HE1))

freq(CAT001-CAT002, φ) = $\gamma \cdot$ (*freq*(WO3-CHI1, φ) + *freq*(NI3-HE1, φ))

internal constructional filler counts

C (*filler*, CAT001-CAT002, c0) = $\gamma \cdot$ (C (*filler*, WO3-CHI1, w0) + C (*filler*, NI3-HE1, n0))

which, due to the narrow type constraints on w0 and n0 effectively means that

C (WO3-N, CAT001-CAT002, c0) = γ C (WO3-N, WO3-CHI1, w0)

C (NI3-N, CAT001-CAT002, c0) = γ C (NI3-N, NI3-HE1, n0)

which works out to be

C (WO3-N, CAT001-CAT002, c0) = γ *freq*(WO3-CHI1)

C (NI3-N, CAT001-CAT002, c0) = γ *freq*(NI3-CHI1)

if there are no pronunciation variants of *wo3* and *ni3*.

external constructional filler counts

C (CAT001-CAT002, ν, β^{ν}) = $\gamma \cdot$ (C (WO3-CHI1, ν, β^{ν}) + C (NI3-HE1, ν, β^{ν}))

locality counts

C ($LOCAL$, CAT001-CAT002, c0)

= $\gamma \cdot$ (C ($LOCAL$, WO3-CHI1, w0) + C ($LOCAL$, NI3-HE1, n0))

C ($EXTRAPOSED$, CAT001-CAT002, c0)

= $\gamma \cdot$ (C ($EXTRAPOSED$, WO3-CHI1, w0) + C ($EXTRAPOSED$, NI3-HE1, n0))

C ($OMITTED$, CAT001-CAT002, c0)

= $\gamma \cdot$ (C ($OMITTED$, WO3-CHI1, w0) + C ($OMITTED$, NI3-HE1, n0))

C ($UNFILLED$, CAT001-CAT002, c0)

= $\gamma \cdot$ (C ($UNFILLED$, WO3-CHI1, w0) + C ($UNFILLED$, NI3-HE1, n0))

likewise for constituent c1 in CAT001-CAT002

Figure 6.2 Example of updating the statistics after a generalization operation between two constructions of equal length.

All the previous utterances that have been analyzed either of WO3-CHI1 and NI3-HE1 are now analyzable by and thus count towards the CAT001-CAT002 construction. To do this, the final constituent mapping is necessary. Constituent c0 (of type CAT001) in CAT001-CAT002 is mapped to constituent w0 (WO3-N) in WO3-CHI1 and constituent n0 (NI3-N) in NI3-HE1. Constituent c1 (CAT002) in CAT001-CAT002 is mapped to constituent c1 (CHI1-V) in WO3-CHI1 and constituent h1 (HE1-V) in NI3-HE1.

The more aggressive generalization between constructions of different lengths leads to a modified update algorithm. The principle that a generalization replaces the specific constructions still holds, but since multiple generalizations are created out of multiple constructions, each general construction combines only the counts of the specific constructions it replaces. Take CAT003-CHI1 and CAT003-CHI1-YAO4 from Figure 4.16 as example. These are created from the generalization between WO3-CHI1 (I eat) and XI1XI1-CHI1-YAO4 (Xi1Xi1 eat medicine), with the new category CAT003 being {WO3, XI1XI1}. CAT003-CHI1 thus gets the discounted counts of WO3-CHI1 and CAT003-CHI1-YAO4 gets the discounted counts of XI1XI1-CHI1-YAO4 such that the constructional unigram frequencies, for example, are updated with

$$freq(\text{CAT003-CHI1}) = \gamma \cdot freq(\text{WO3-CHI1})$$

$$freq(\text{CAT003-CHI1-YAO4}) = \gamma\, freq(\text{XI1XI1-CHI1-YAO4})$$

and the internal constructional filler counts are updated with

$$C\,(filler, \text{CAT003-CHI1, c0}) = \gamma\, C\,(filler, \text{WO3-CHI1, w0})$$

$$C\,(filler, \text{CAT003-CHI1, c1}) = \gamma\, C\,(filler, \text{WO3-CHI1, c1})$$

$$C\,(filler, \text{CAT003-CHI1-YAO4, c0}) = \gamma\, C(filler, \text{XI1XI1-CHI1-YAO4, x0})$$

$$C\,(filler, \text{CAT003-CHI1-YAO4, c1}) = \gamma\, C\,(filler, \text{XI1XI1-CHI1-YAO4, c1})$$

$$C\,(filler, \text{CAT003-CHI1-YAO4, y2}) = \gamma\, C\,(filler, \text{XI1XI1-CHI1-YAO4, y2})\ .$$

Construction revision

A revision is the modification of meaning constraints or the addition of syntactic cues to an existing construction to resolve its conflict with another construction. The revised construction is intended to replace the original but they are put into competition since the learner cannot be sure of the correctness of the revision (lack of semantics in function words as well as inconsistency in the input come to mind). The revised construction receives all the counts of the original discounted by the factor γ. In other words, the constructional unigram frequency of the revised construction, $\mathrm{freq}(\alpha_{\mathrm{new}})$, is the discounted constructional unigram frequency of the original construction, $\gamma\mathrm{freq}(\alpha)$. The constructional bigram frequency of the new construction $\mathrm{freq}(\varphi, \alpha_{new})$ is the discounted bigram frequency of the original constructions, $\gamma \sum_i \mathrm{freq}(\varphi, \alpha)$ for all constructions φ. The same applies for the other bigram frequency $\mathrm{freq}(\alpha, \varphi)$.

The internal constructional filler counts for constituents that exist in the original construction are discounted and re-used in the revised construction, and for the newly added constituents the internal constructional filler count is the same as the constructional unigram frequency. That is, $\mathrm{C}\left(filler_\beta, \alpha_{new}, \beta_j^{\alpha_{new}}\right) = \gamma \cdot \mathrm{C}\left(filler_\beta, \alpha, \beta_j^{\alpha}\right)$ if $\beta_j^{\alpha_{new}}$ is also in α and $\mathrm{C}\left(filler_\beta, \alpha_{new}, \beta_j^{\alpha_{new}}\right) = \gamma \cdot \mathrm{freq}(\alpha)$ otherwise. The external constructional filler count of construction α_{new} in constituent β_k^{υ} in user υ, $\mathrm{C}(\alpha_{new}, \upsilon, \beta_k^{\upsilon})$, is a straightforward discount of the external constructional filler count of original construction, $\gamma\mathrm{C}(\alpha_{\mathrm{i}}, \upsilon, \beta_k^{\upsilon})$. The locality counts for constituents that exist in the original construction are discounted and re-used in the revised construction, and *LOCAL* count for the newly added constituents is the same as the constructional unigram frequency. That is, $\mathrm{C}\left(locality_\beta, \alpha_{new}, \beta_j^{\alpha_{new}}\right) = \gamma\mathrm{C}\left(locality_\beta, \alpha, \beta_j^{\alpha}\right)$ if $\beta_j^{\alpha_{new}}$ is in α and $\mathrm{C}\left(locality_\beta, \alpha_{new}, \beta_j^{\alpha_{new}}\right) = \gamma\mathrm{freq}(\alpha)$ otherwise.

The example used here is BA3-YUE4LIANG4-CAT006 (CV_{obj} moon tear/rip/damage) from Figure 5.5, which was revised from YUE4LIANG4-CAT006 (moon tear/rip/damage) with an additional coverb BA3 which marks the direct object.

REVISION
new construction: BA3-YUE4LIANG4-CAT006 (CV_{obj} moon tear/rip/damage)
revised from YUE4LIANG4-CAT006 (moon tear/rip/damage)

constructional unigram frequency
freq(BA3-YUE4LIANG4-CAT006) = $\gamma \cdot$ *freq*(YUE4LIANG4-CAT006)

constructional bigram frequency
freq(φ, BA3-YUE4LIANG4-CAT006) = $\gamma \cdot$ *freq*(φ, YUE4LIANG4-CAT006)
freq(BA3-YUE4LIANG4-CAT006, φ) = $\gamma \cdot$ *freq*(YUE4LIANG4-CAT006, φ)

internal constructional filler counts
C (BA3, BA3-YUE4LIANG4-CAT006, b0) = γ *freq*(YUE4LIANG4-CAT006)
C (*filler*, BA3-YUE4LIANG4-CAT006, y1) = γ C (*filler*, YUE4LIANG4-CAT006, y0)
C (*filler*, BA3-YUE4LIANG4-CAT006, c2) = γ C (*filler*, YUE4LIANG4-CAT006, c1)

external constructional filler counts
C(BA3-YUE4LIANG4-CAT006, ν, β^{ν}) = γ C(YUE4LIANG4-CAT006, ν, β^{ν})

locality counts
C ($LOCAL$, BA3-YUE4LIANG4-CAT006, b0) = γ *freq*(YUE4LIANG4-CAT006)
C (*locality*, BA3-YUE4LIANG4-CAT006, y1) = γ C (*locality*, YUE4LIANG4-CAT006, y0)
C (*locality*, BA3-YUE4LIANG4-CAT006, c2) = γ C (*locality*, YUE4LIANG4-CAT006, c1)

Figure 6.3 Example of updating the statistics after a revision operation.

Constituent omission

A constructional constituent is learned to be omissible or optional through contrasting two constructions that differ by one or more constituents in length (e.g. GEI3-CLAUSE and GEI3-WO3 introduced from Figure 5.2). Let s be the shorter construction and l be the longer construction. Constituents in l but not in s (e.g. w1 in GEI3-WO3) have their omission /optionality statuses modified so that l can now be used to analyze sentences originally covered by s. In this sense, a constituent omission operation is a special case of the concrete generalization operation and the statistics update for the resulting modified longer construction l' proceeds quite like that for generalization. However, in the current implementation, s is removed from the grammar at the end of this operation and l' is kept in place of l, so no discounting is necessary.

Concretely, the constructional unigram frequency of the modified longer construction $\text{freq}(l')$ is the sum of the constructional unigram frequencies of the shorter construction and its original version, i.e., $\text{freq}(s) + \text{freq}(l)$. The constructional bigram frequencies are summed in the same way for all constructions φ in the grammar such that $\text{freq}(\varphi, l') = \text{freq}(\varphi, \text{s}) + \text{freq}(\varphi, l)$ and $\text{freq}(l', \varphi) = \text{freq}(s, \varphi) + \text{freq}(l, \varphi)$.

The external constructional filler counts are again a straightforward summation of the counts held by the original pair of constructions, i.e. $\text{C}(l', \upsilon, \beta_k^{\upsilon}) = \text{C}(s, \upsilon, \beta_k^{\upsilon}) + \text{C}(l, \upsilon, \beta_k^{\upsilon})$. The update of the internal constructional filler counts and locality counts requires a bit of care and the example will make this a lot clearer. For the shared constituents between s and l, both the constructional filler counts and the locality counts are summed as usual, i.e. $\text{C}\big(filler_{\beta}, l', \beta_j^{l'}\big) = \text{C}\big(filler_{\beta}, s, \beta_j^{s}\big) + \text{C}\big(filler_{\beta}, l, \beta_j^{l}\big)$ and $\text{C}\big(locality_{\beta}, l'\big) = \text{C}\big(locality_{\beta}, s\big) + \text{C}\big(locality_{\beta}, l\big)$.

The differing constituents are the ones that are now omissible or optional, depending on whether they are connected semantically to core roles of the meaning of the construction. The constructional filler counts of these constituents, which are already a part of l, are retained. The locality probability is the most important aspect of this learning operation. The $LOCAL$ count of constituent β_j^l is the number of times the longer construction is used. If the constituent is omissible, the number of times the shorter construction is the *omitted* count. If the constituent is optional, the unigram frequency of the shorter construction is the *unfilled* count.

$$C(local_\beta, l) = \text{freq}\,(l)$$

$$C(omitted_\beta, l) = \text{freq}(s) \text{ if } \beta_j^l \text{ is omissible}$$

$$C(unfilled_\beta, l) = \text{freq}(s) \text{ otherwise}$$

Returning to the GEI3-CLAUSE and GEI3-WO3 introduced in Figure 5.2, which led to the omissible constituent w0 in the longer construction GEI3-[WO3], the following figure shows how the update is performed so that the omission probability of constituent w0 is properly calculated.

CONSTITUENT OMISSION
new construction: GEI3-[WO3] (give [me])
original constructions: GEI3-CLAUSE and GEI3-WO3

constructional unigram frequency
freq(GEI3-[WO3]) = *freq*(GEI3-CLAUSE) + *freq*(GEI3-WO3)

constructional bigram frequency
freq(φ, GEI3-[WO3]) = *freq*(φ, GEI3-CLAUSE) + *freq*(φ, GEI3-WO3)
freq(GEI3-[WO3] , φ) = *freq*(GEI3-CLAUSE, φ) + *freq*(GEI3-WO3, φ)

internal constructional filler counts
C (*filler*, GEI3-[WO3], g0) = C (*filler*, GEI3-CLAUSE, g0) + C (*filler*, GEI3-WO3, g0)
C (*filler*, GEI3-[WO3], w1) = C (*filler*, GEI3-WO3, w1)

external constructional filler counts
C(GEI3-[WO3], v, β^v) = C (GEI3-CLAUSE, v, β^v) + C (GEI3-WO3, v, β^v)

locality counts
C (*locality*, GEI3-[WO3], g0) = C (*locality*, GEI3-WO3, g0)
C (*LOCAL*, GEI3-[WO3], w1) = *freq*(GEI3-WO3)
C (*OMITTED*, GEI3-[WO3], w1) = *freq*(GEI3-CLAUSE)

Figure 6.4 Example of updating the statistics after a constituent omission operation.

Category merge

A category merge merges an existing category α_2 into α_1, i.e. all the subtypes of α_2 are made subcases of α_1 and α_2 is subsequently removed. Because all the grammar statistics are tracked on concrete constructions, none of the statistics need to be udpated.

Category expansion

Finally, constructional category membership can be extended to constructions that are previously non-members. For example, CAT033 from Figure 5.10 is extended from CUO1-V (mash), RENG1-V (throw), and TI1-V (kick) to other constructions with a meaning of agentive impact, such as DA3-V (beat) and PAI1-V (slap). Users of the expanded category can have their constructional filler probabilities altered to give counts to the new category members, but this is imprudent since the category expansion is an aggressive operation to begin with. If new category members are also given additional constructional filler counts, the sudden influx of category members may dilute the probability mass so much from the existing members that the probability of using any one of them (and therefore the probability of the construction) may become too low to compete with other constructions. The model thus chooses not to adjust these counts, leaving the probability mass to be distributed to the new category members in the smoothing process.

Decay

When a construction is removed from the grammar due to decay, its corresponding entries in the statistics are simply removed and the distributions re-normalized.

Summary

Figure 6.5 shows a summary of the update strategy for each statistic after each learning operation. Again, the guiding principle behind all these udpate heuristics is to pretend as if the learner has perfect knowledge in creating new constructions and assigning occurrence counts using all the previously encountered data as support. A discount factor $0 \leq \gamma \leq 1$ is then applied to these counts to adjust for uncertainty. Crucially, these updates can be done incrementally based on existing counts, and no re-estimation using the entire corpus is necessary.

	Compositio n	Concrete Generalization (across $\alpha_{1..i}$)	Revision (α_{new} based on α)	Omission (based on s and l)	Category Merge (of α_1 and α_2)	Category Expansion (of α_{new})
$freq(\alpha_{new})$	$\gamma \cdot 1$	$\gamma \sum_i freq(\alpha_i)$	$\gamma freq(\alpha)$	$freq(s) + freq(l)$	N/A	N/A
$freq(\varphi, \alpha_{new})$	N/C	$\gamma \sum_i freq(\varphi, \alpha_i)$	$\gamma freq(\varphi, \alpha)$	$freq(\varphi, s) + freq(\varphi, l)$	N/A	N/A
$freq(\alpha_{new}, \varphi)$	N/C	$\gamma \sum_i freq(\alpha_i, \varphi)$	$\gamma freq(\alpha, \varphi)$	$freq(s, \varphi) + freq(l, \varphi)$	N/A	N/A
$C(filler_\beta, \alpha_{new}, \beta_j^{\alpha})$	$\gamma \cdot 1$	$\gamma \sum_i C(filler_\beta, \alpha_i, \beta_j^{\alpha_i})$	$\gamma C(filler_\beta, \alpha_2, \beta_j^{\alpha})$ if β_j^{α} exists	$C(filler_\beta, s, \beta_j^{s}) + C(filler_\beta, l, \beta_j^{l})$ if β_j^{l} and β_j^{s} both exist, unchanged o.w.	N/A	N/A
$C(\alpha_{new}, \upsilon, \beta_k^{\upsilon})$	N/C	$\gamma \sum_i C(\alpha_i, \upsilon, \beta_k^{\upsilon})$	$\gamma C(\alpha, \upsilon, \beta_k^{\upsilon})$	$C(s, \upsilon, \beta_k^{\upsilon}) + C(l, \upsilon, \beta_k^{\upsilon})$	N/A	N/C
$C(locality_\beta, \alpha_{new}, \beta)$	$LOCAL\ \gamma$	$\gamma \sum_i C(locality_\beta, \alpha_i, \beta)$	$\gamma C(locality_\beta, \alpha)$	$C(locality_\beta, s) + C(locality_\beta, l)$ if βi in both s and l, o.w. $C(OMITTED, l, \beta) = freq(s)$ $C(LOCAL, l, \beta) = freq(l)$	N/A	N/A

Figure 6.5 Strategy adopted by the learner in updating grammar statistics during learning. These include the unigram and bigram frequencies, internal constructional filler counts, external constructional filler counts, and the constituent locality counts.

6.3 Calculating the grammar statistics

The unigrams and bigrams matter to the learner in only two ways: unigram frequencies are used to update the locality probabilities after an omission operation, and bigrams are used to filter potential revisions (i.e., recruiting frequently collocated function words as new constituents). Neither of these statistics is used by the analyzer and is therefore not smoothed.

The constructional filler probabilities are obtained by smoothing and $\mathrm{P}(\mathit{filler}_\beta \,|\, \mathit{expressed}_\beta, \mathit{typeConstraint}_\beta)$ normalizing the constructional filler counts. The smoothing function is a linear combination of $\mathrm{P}(\mathit{filler}_\beta \,|\, \mathit{expressed}_\beta, \alpha)$ and its backoffs given by:

$$\begin{aligned}\mathrm{P}_{\mathit{smoothed}}(\mathit{filler}_\beta \,|\, \mathit{expressed}_\beta, \alpha) \;=\; & \rho \mathrm{P}(\mathit{filler}_\beta \,|\, \mathit{expressed}_\beta, \alpha) \\ & +\sigma(1-\rho)\mathrm{P}(\mathit{filler}_\beta \,|\, \mathit{expressed}_\beta, \mathit{typeConstraint}_\beta) \\ & +(1-\sigma)\mathrm{P}(\mathit{filler}_\beta)\end{aligned}$$

where is the context-free backoff that is conditioned only on the type constraint of the constituent, and $\mathrm{P}(\mathit{filler}_\beta)$ is the uniform backoff of all type-suitable fillers. The held-out mass, captured by the constants ρ and σ, are given by the standard Witten-Bell smoothing procedure.

The locality probabilities are calculated from the locality counts without any smoothing.

Chapter 7.

Mandarin Chinese learning experiments

The learning model is tested on naturalistic child language data as well as an artificial language. This chapter describes the computational experiments with Mandarin Chinese data and the next chapter describes those with a miniature language, but the procedure for both tasks is the same: the learning model is set up with an initial grammar and learning is performed on a training set. At set intervals, the learned grammar is tested on a validation set. Note that this validation procedure does not provide the learner with any feedback; it only serves as an indicator of how well the learning model is performing.

7.1 Learning data

The data has been partly described in Section 2.5; an augmented set is used here. The training set consists of 150 short dialogues and the validation set consists of 4 long dialogues, all of which are taken from the Tardif Beijing corpus in CHILDES (MacWhinney, 2000; Tardif, 1993; 1996). This is a longitudinal corpus of naturalistic parent-child interaction. The short dialogues are created from the transcripts of subjects CXX (Xi Xi), HY (Hao Yu), and LXB (Xiao Bing) while the long dialogues are taken from a forth subject WW (Wei Wei). Those four children range from 1;9.3 to 1;10.28 at the start of the study and range from 2;1.4 to 2;3.2 at the end of the fifth recording sessions.

The short training dialogues are sections of the transcripts that are manually selected to maintain topic coherence, containing generally no more than two ongoing activities though the number of referenced objects varies. Unintelligible sentences are removed but the selected utterances are close together temporally. The long testing dialogues, on the other hand, are contiguous segments of transcripts taken from the beginning and the end of two recording sessions. Few manual edits were made except to remove entire utterances that are unintelligible.

Transcribed utterances are split into separate utterances wherever pauses are notated, as run-on sentences containing different speech acts are fairly common. The resulting training corpus contains 2071 utterances of which 1637 are child-directed. The resulting validation corpus contains 317 utterances of which 229 are child-directed. Each short dialogue contains on average 13.8 utterances and each long dialogue contains on average 79.3 utterances. For the training corpus, the mean length of the parental utterances (parental MLU) is 3.59 and the meaning length of the child utterances (child MLU) is1.79. For the validation corpus, the mean length of parental utterances is 3.01 and that of the child utterances is 2.68. Based on the number of open class nouns and verbs needed in the starting lexicon to process the combined corpus, there are roughly 252 verb types[31] and 178 noun types present in the utterances.

Both the training and validation sets are annotated with event and speech acts. The validation set is additionally annotated with gold standard semantic annotations. Conveniently, due to the parsing experiment described in Chapter 2, the first 35 short dialogues in the training set are also annotated with gold standard semantic annotation and can be used as a secondary validation set. The secondary validation set (henceforth called the seen validation data) serves as a sort of sanity check for the learning model while the long dialogues (henceforth called the unseen validation set) truly tests the learner's ability to generalize from the training data. The event

[31] including stative verbs such as *hao3+kan4* (pretty), which in English will be more akin to adjectives. In Mandarin no copula verb is necessary and therefore these are often analyzed as stative verbs in the linguistic literature. These account for about 58 of the 252 verb types.

annotation, speech act annotation and gold standard annotation have been briefly described in Section 2.5 but will be expanded up on here for completeness. For those interested in the details, the annotated data and both the learner grammar and handwritten grammar (from Chapter 2) are made available on the ECG wiki (http://ecgweb.pbwiki.com).

		Short 0 - 35	**Short 35-150**	**Long 0-4**
	purpose	Training Set / Seen Validation	Training Set	Unseen Validation
corpus properties	# of utterances	385	1686	317
	# child-directed	318	1319	229
	MLU (parental)	3.72	3.56	3.01
	MLU (child)	1.64	1.82	2.68
annotation	event	yes	yes	yes
	speech act	yes	yes	yes
	goldstandard	yes	no	yes

Figure 7.1 Summary of the corpus data used in the Mandarin Chinese grammar learning task

Event annotation

The attentional focus of the learner is continuously estimated using the context model, which relies on data annotations for updates. The setting and event annotation described in Section 2.5 is carried out on the expanded training data. Event categories drawn from the ontology as well as event participants are annotated in the dialogue wherever an event is presumed to have occurred, at the annotator's discretion. Each dialogue is thus situated in a reasonable scene. Figure 7.2 illustrates typical data annotation with another dialogue.

<event cat="Find" id="find01">
<binding field="finder" ref="MOT"/>
<binding field="target" ref="crayon"/>
</event>

MOT: *zher4 ne*
(here SFP / here it is)

MOT: *ma1 gei3 ni3 na2*
(mother for you retrieve / mother's getting it for you)

<event cat="Fetch" id="fetch02">
<binding field="fetcher" ref="MOT"/>
<binding field="fetched" ref="crayon"/>
</event>

MOT: *gei3 ni3 na2 cai3+bi3*
(for you retrieve crayon / I'm getting you a crayon)

<event cat="Give" id="give03">
<binding field="giver" ref="MOT"/>
<binding field="recipient" ref="CHI"/>
<binding field="theme" ref="crayon"/>
</event>

MOT: *ni3 rang4 a1+yi2 kan4+kan4 ni3 de bi3 hao3+kan4 bu4 hao3+kan4*
(you let aunt look you ASSOC pen pretty NEG pretty /
you let aunt see whether your crayon is pretty)

CHI: *a* (INJ)

Figure 7.2 An example scene where the mother searchers for and retrieved a crayon for the child.

Speech act annotation

Each utterance is annotated with discourse information, including the current speaker, addressee, intonational forcefulness and speech act, which is one of the cues used in the context-fitting heuristics (see Section 2.4). Dore suggests that primitive speech acts expressed by children at the one-word stage include *labeling*, *repeating*, *answering*, *requesting (action)*, *requesting (answer)*, *calling*, *greeting*, *protesting*, and *practicing* (Dore, 1974). For the purpose of the learning model, this list is adapted to create a list of primitive speech acts that seem reasonable for a child of grammar-learning age, including ones that do not elicit a response (explaining, answering, approving, admonishing), ones that do elicit a response (requesting action, requesting answer, calling) and ones that does not necessarily require an audience (exclaiming, practicing). Figure 7.3 gives a more detailed description of each of the speech act types.

Explaining	Describing an object or event. The object is usually present in the environment, while the event can be co-timed, immediately before or immediately after the utterance.
Answering	Answering a question or request.
Approving	Usually performed by the parent, expressing approval of a prior child utterance or action.
Admonishing	Usually performed by the parent, criticizing the child for a prior event, preventing the child from performing future actions, or threatening with punishment.
Requesting action	Making a demand that requires a physical response from the addressee, often accompanied by gesture. This includes requests for the child to perform an action and requests for parents to retrieve objects.
Requesting answer	Asking a question that requires a verbal response from the addressee.
Calling	Beckoning the addressee.
Exclaiming	Producing exclamations that express surprise, joy, or pain.
Practicing	Usually done by the child, producing words or phrases that are either incoherent or do not pertain to the current interaction.

Figure 7.3 List of speech acts used to annotate both adult and child utterances in the corpus.

Initial grammar

In order to provide coverage for the training and validation data, the initial grammar consists of roughly 300 schemas and 706 constructions. The schemas are made up of:

- 11 structural schemas such as EVENT_DESCRIPTOR and RD
- 84 "closed-class" conceptual schemas that define the process lattice, image schemas, speech acts, etc
- 205 "open-class" process schemas

Included in the set of constructions are:

- primary abstract constructions: CLAUSE, PHRASE, MORPHEME, WORD, NUMBER, DIGIT, INTERJECTION, INTERJECTION_MORPHEME
- 81 closed class lexical constructions (with meaning), such as pronouns, negation words, path particles, directional particles, and digits
- 252 open class verbs and 178 open class nouns [32]
- 89 function word forms (without meaning), such as coverbs, classifiers, locative particles, adverbials and conjunctions

The ontology contains an additional 204 entity types and about 190 simulation scripts corresponding to the annotated events are used to update the context model.

7.2 Experiment 1: Mandarin Chinese CHILDES corpus — basic experiment

Training procedure

The basic learning experiment was run with all the learning operations enabled except decay. This is done to make tracing through the learning and generalization history easier. The statistic update discount factor γ was set to 1, 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 -10

[32] Often the same word is pronounced in different tones in fluent speech and transcribed as such. Pronunciation variants of a word are represented as subcases of an abstract construction.

(in log probability scale). The learner then fits each analysis to context. The analysis with the best contextual fit (using an internal metric based on number of referenced entities as a proxy for coherence) is selected for learning.

The learner attempted to perform up to 4 learning iterations over the entire 150 short dialogues (or over 2000 total utterances) in the training corpus, but without a decay mechanism, the grammar quickly grew too large and the analyzer ran out of memory during one of its validation sequences after 750 learning episodes[33]. The results reported here were thus based on barely over one-third of the training data available. Regardless, some 18 categories and 515 concrete phrasal and clausal constructions were learned at the end of the experiment. The next subsections describe these results in qualitative and quantitative terms.

There is considerable current work in the research group by John Bryant and others to scale and optimize the analyzer for larger grammars. For the current work, a few other combinations of learning operations were tested out and will be described in Section 7.3.

Qualitative results

To give a concrete sense of what the learning model does with the input data, the next three figures show excerpts of the learning steps taken by the model at different stages of the learning process. For brevity only the construction name and a summary is shown; the notation is explained in the caption of Figure 7.4. With the belief that mistakes are often more informative than the correct output, this section tries to give as much coverage about bad choices that the model made as the good ones.

[33] Because of the many small construction fragments available, the analyzer sometimes get horrendously garden-pathed and goes off on a memory-intensive search over all the possible ways to try and connect the words in the utterance under the same root in the analysis.

As expected, early constructions learned by the model are very lexically specific. Highly retricted grammatical categories are first created and they got progressively bigger. The earliest generalizations tend to have only one constituent that allows an abstract construction as its type constraint (i.e. only one open "slot" such as e.g. ___ *eat*). This comes about not by specific design of the model but as a result of how the model incrementally creates generalizations as they become available. The first construction with two open slots appear after 47 learning episodes and is a limited construction that expresses {you, mother} - {apply, put}.[34] These kinds of constructions get more and more common as further generalizations are made, and constructions with omissible and optional constituents were eventually learned as well.

In the earliest few episodes (a - c) of Figure 7.4, the learner composed lexically specific constructions with very strict contextual restrictions on both the discourse role and the ontological type of the semantic arguments, shown in angle brackets. After encountering a few uses of *mo3* (apply), some negated and some not, the learner decided that the negation particle BIE2-F (don't) is an optional constituent of the BIE2-MO3-c005 construction (d).

[34] apply as in the applying lotion sense of the word, or moving some substance from a source to some surface.

Operation		
	Resulting Construction	**Meaning Gloss + Contextual Restriction**
a)	Compose LIANG4-V (switch on) and DENG1-N (light)	
	LIANG4-DENG1-c002	<addressee#child>- SWITCH_ON - desklamp
b)	Compose HAO3WANR2-V (amusing)	
	HAO3WANR2-c003	< attnFocus#desklamp> - AMUSING
c)	Compose BIE2-F (don't) and MO3-V (apply)	
	BIE2-MO3-c005	<addressee#child > - NEG - APPLY - attnFocus#lotion>
d)	Optionalize b0 in BIE2-MO3-c005 (don't apply)	
	[BIE2]-MO3-c005	[optional NEG] – APPLY

Figure 7.4 The earliest learning operations carried out by the model and the resulting constructions. Glosses for lexical items are provided also in parenthesis when appropriate.

A guide to reading this short-hand: The names of lexical constructions end in -N (noun) / -V (verb) / -F (function words). This notation is only to aid the reader and is not meaningful to the model. The names of learned construction ends in -c followed by a 3-digit ID.

On the right, the meaning poles of the constructions are shown in ALL CAPS such as SWITCH_ON. The arguments to the construction are shown in an English-centric order, so that <addressee#child>- SWITCH_ON - desklamp denotes a SWITCH_ON event in which the child is the agent and the desklamp is the patient. A contextual constraint on a core argument that is not present constructionally is shown in angle brackets; the hash mark separates the discourse role restriction from the contextual type constraint, as in <addressee#child>. An empty angle bracket means that there are no restrictions. Omissible constituents are denoted in square brackets and optional constituents are marked as such inside square brackets.

Notations in the next diagrams: Members of constructional categories are put in curly brackets, as in CAT021: {HAO3WANR2-V, HUAI4-V}. Speech act restrictions are notated in parenthesis only if they make an important distinction in the example.

e) Compose HUAI4-V (broken)

HUAI4-c020	< attnFocus#car> - BROKEN (explaining)

f) Generalize HAO3WANR2-c003 (amusing) and HUAI4-c020 (broken)

CAT021: {HAO3WANR2-V (amusing), HUAI4-V (broken)}

CAT021-c022	<attnFocus#ManipulableObject> - INTRANSITIVE_STATE

g) Generalize XING2-c010 (good-enough) and CHE1_HUAI4-c024 (car broken)

CAT025: {XING2-V (good-enough), HUAI4-V (car broken)}

CAT025-c026	<addressee#child > - GOOD_ENOUGH
CHE1-CAT025-c027	car - BROKEN (explaining)

h) Generalize HUAI4-c020 (broken) and CHE1-CAT025-c027 (car {good-enough, broken})

CAT025-c029	< attnFocus #car> - BROKEN (explaining)

i) Compose CHE1-N (car) and CAT025-c029 ({good-enough, broken})

CHE1-CAT025-c031	car - BROKEN (approving)

j) Generalize CAT021-c022 and CHE1_CAT025-c031 (car {good-enough, broken})

CAT032: {CAT021 , CAT025}

CAT032-c033	< attnFocus #ManipulableObject> - INTRANSITIVE_STATE
CHE1-CAT032-c034	car - BROKEN (approving)

k) Merge CAT021 and CAT025 into CAT032

CAT032: { HAO3WANR2-V (amusing), XING2-V (good-enough), HUAI4-V (broken) }

l) Expand CAT032 onto other words with stative meaning

CAT032: { HAO3WANR2-V (amusing), XING2-V (good-enough), HUAI4-V (broken), WAN2-V (finished), XIANG1-V (fragrant), GAN1JING4-V (clean), GOU4-V (enough), …}

Figure 7.5 After a number of lexically-specific constructions are learned, generalizations start to appear.

Operation

Resulting Construction — **Meaning Gloss + Contextual Restriction**

m) Generalize A1YI2-CANG2-CXN697 (aunt hide) and TA1-CAT177-C545

- (existing) CAT045: {NI3_VARIANT-N (you), ZAN2MEN-N (we), BAO3BAO3-N (child), MA1MA_VARIANT-N (mother), NAI3NAI-N (grandma), ...}
- (existing) CAT177: {PAO3-V (run), ZOU3-V (go), LAI2-V (come), CHU1-V (exit), SHANG4-V (ascend), ...}

CAT045-CAT177-c713 — < > - TRANSLATIONAL_SELF_MOTION - <solid>

n) Revise CAT045-CAT177-c713 adding constituent NA2-V (bring)

CAT045-NA2-CAT177-c841 — < > - TRANSLATIONAL_SELF_MOTION - <solid>

...

o) Omit constituent n1 in CAT177-NEI4-c676

CAT177-[NEI4]-c676 — TRANSLATIONAL_SELF_MOTION – [there]

...

p) Generalize CAT981-CAT419-c922 and NA2-CAI3BI3-c1033

- (existing) CAT981: {CA1CA1-V (wipe), DING3-V (throw), DENG4-V (kick), DONG4-V (move), JIA1-V (pick), ...}
- (existing) CAT419: {XIONG2-N (bear), QIANG1-N (toy gun), QIN2-N (piano), TU3DOU4-N (potato), BAN3DENGR4-N (stool), ...}

CAT981-CAT419-c1035 — <addressee#child> - GRASP - ManipulableObject (no speechact restrictions)

q) Generalize CAT981-CAT419-c1035 and qi2_mo2tuo1che1-Cxn1831

CAT981-CAT419-c1834 — <addressee#child> - CONTINUOUS_FORCE_APPLICATION - ManipulableObject (no speech act restriction)

Figure 7.6 A sample of the later learning operations taken by the learner and the resulting constructions.

Once a number of lexically-specific constructions had been composed and used, the learner was able to form generalizations. From these first few generalizations in Figure 7.5 it is immediately apparent that the model is very conservative and takes small, incremental steps in its learning. One of the first generalization the learner made (f) was between two clausal constructions with stative meaning: HAO3WANR2-c003 (amusing) and HUAI4-c020 (broken), resulting in a category CAT021 over the two verbs HAO3WANR2-V (amusing) and HUAI4-V (broken). The next generalization (g) was between two stative constructions of different lengths: XING2-c010 (good-enough) and CHE1_HUAI4-c024 (car broken). This resulted in another category CAT025 over the verbs XING2-V (good-enough) and HUAI4-V (broken). It also created a generalized version of each of the specific constructions. The general constructions retained the meaning pole restrictions of the specific versions, as shown in Figure 7.7, which is arguably not very useful:

```
Construction XING2-c010
    subcase of CLAUSE
    constructional
        constituents
            x0:XING2-V
    meaning:GOOD_ENOUGH
        evokes RD as rd0
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> x0.m
            rd0.referent <--> x0.m.protagonist
            rd0.referent <--> rd0.ontological_category
            rd0.ontological_category <-- @Child
            rd0.discourse__role <-- @Addressee
            DS.speech_act <-- Requesting_Action
```

```
Construction CAT025-c026
    subcase of CLAUSE
    constructional
        constituents
            c0:CAT025
    meaning:GOOD_ENOUGH

            ⋮
```

Figure 7.7 A limited generalization between constructions of different lengths results in the CAT025-c026 construction on the right, which is only slightly more general than XING2-c010 in its constituent type requirements (in bold italics) but retain exactly the same semantic restrictions on the overall event type.

The learning algorithm allows generalizations between different omission patterns so that such data as (the equivalent of) *I eat* and *you eat rice* can lead to the generalizations: {I, you}-eat and {I, you}-eat-rice. In not wanting to presuppose a verb bias, the current algorithm makes no special provisions for the cases when the verbs differ. The learning behavior is therefore such that the semantic restrictions from the specific construction (such as GOOD_ENOUGH in XING2-c010 above) get retained in the more general version, in this case effectively limiting CAT025-c026 to only work with the verb XING2-V. [35]

The category CAT025 in (g) overlaps with CAT021 in (f) but they were not merged until later on in (j), when the need presented itself through another generalization. Since CAT021 and CAT025 were made subcases of CAT032 in that generalization, a category merge was automatically triggered (k). CAT021 and CAT025 were removed from the grammar and their category members, HAO3WANR2-V (amusing), XING2-V (good-enough), HUAI4-V (broken), were made direct subcases of CAT032. With three members in this category, the learner took a leap of faith and extended this category to all other words with a stative process meaning in (l), including words that had not been encountered in the learning input, such as WAN2-V (finished), XIANG1-V (fragrant), and GAN1JING4-V (clean). Although some existing constructions such as CHE1-CAT032-c034 were very semantically specific and were not able to take advantage of this newly expanded category, the CAT032-c033 construction was now able to cover any sentence that used one intransitive verb to describe any manipulable object.

Notice how a number of compositions were made in the interim between (g) and (k) because the utterances encountered differed in the speech act restrictions from existing constructions (see (e) and (i)). Differing constructions of this sort were eventually merged further down the line, as is made evident by the categories and concrete constructions present in the grammar much later in the learning process , shown in Figure 7.6.

[35] Obviously, no ends of learning optimizations are possible here, but the goal of this dissertation is to create a set of reasonably simplistic learning operations to tease out what is important to the cognitive task of grammar learning rather than an engineering system that does the job in the best possible manner.

The learning operations shown in Figure 7.6 are those carried out after at least 300 learning episodes, about halfway through the current learning experiment. A number of rather well-established categories already existed prior to the generalization in (m) between A1YI2-CANG2-CXN697 (aunt hide) and TA1-CAT177-C545 (he {go, run, exit, ...}). The two relevant ones are shown in Figure 7.8: CAT045 consists of human-denoting words, and CAT177 of TRANSLATIONAL_SELF_MOTION words. When the generalization (m) was carried out, these existing categories were conveniently used in the new construction, CAT045-CAT177-c713. This construction is itself fairly unrestrictive in meaning. It states that the mover, filled in by something of CAT045 but otherwise unrestricted in its ontological type, moves towards a goal that is a solid object in context. The construction in its entirety is shown in Figure 7.9.

Soon after learning the CAT045-CAT177-c713 construction, the learner found that this was in conflict with another existing construction, CAT419-CAT177-c420. CAT419 is a category of words referring to manipulable objects, and somewhere along the line the learner learned that a noun preceding a motion verb can denote the goal of the motion. That is only half the story in Mandarin Chinese: the destination of motion can be mentioned before the verb, but a coverb such as *wang3* is necessary. The learner was correct in believing that one of CAT045-CAT177-c713 and CAT419-CAT177-c420 required revision, but it chose to revise the wrong construction (partly because it encountered data that used the former construction next).

```
Abstract Construction CAT045
  subcase of MORPHEME
  meaning:@Human
    evokes RD as rd
    constraints
      rd.referent <--> self.m
      rd.ontological_category<-->self.m
```

```
Abstract Construction CAT177
  subcase of MORPHEME
  meaning:TRANSLATIONAL_SELF_MOTION
```

Figure 7.8 Two grammatical categories in the grammar half-way through the learning experiment. CAT045 **is a category of words that refer to humans.** CAT177 **is a category of words that have a meaning of** TRANSLATIONAL_SELF_MOTION**.**

```
Construction CAT045-CAT177-c713
    subcase of CLAUSE
    constructional
        constituents
            c0:Cat045
            c2:Cat177
    form
        constraints
            c0.f before c1.f
    meaning:TRANSLATIONAL_SELF_MOTION
        evokes RD as rd0
        evokes RD as rd1
        evokes DISCOURSE_SEGMENT as DS
        constraints
            self.m <--> c1.m
            c1.m.mover <--> c0.m
            rd0.referent <--> c1.m.mover
            rd0 <--> c0.rd
            rd1.referent <--> c1.m.goal
            rd1.referent <--> rd1.ontological_category
            rd1.ontological_category <-- @Solid
            DS.speech_act <-- EXPLAINING
```

Figure 7.9 A generalization between A1YI2-CANG2-CXN697 **(aunt hide) and** TA1-CAT177-C545 **(he {go, run, exit, ..}) creates a new construction that uses the existing categories** CAT045 **and** CAT177**.**

```
Construction CAT045-NA2-CAT177-c841
    subcase of CLAUSE
    constructional
        constituents
            c0:Cat045              [0.71, 1.0]
            optional n1: na2-V     [0.58]
            c2:Cat177
    form
        constraints
            c0.f meets n1.f
            n1.f meets c2.f
    meaning:TRANSLATIONAL_SELF_MOTION

                         ⋮
```

Figure 7.10 A misguided but understandable attempt to revise CAT045-CAT177-c713 **by inserting another verb** NA2-V **(bring) into the construction. The learner mistakenly believes that the verbs in** CAT177 **co-occur with** NA2-V **because bleached forms of most of those verbs act as path particles in Mandarin Chinese.**

This mistake was further compounded by the fact that a good number of translation motion verbs in Mandarin Chinese such as *zou3* (go away), *chu1* (exit), *qu4* (go), *lai* (come) have a bleached form that are used as path particles. These path particles, unsurprisingly, collocate with other forceful motion verbs. Both senses of the word exist in the initial grammar (a topic we will return to in the final chapter), but these particles are often misanalyzed. Here in (n), then, the learner decided to add an additional constituent `NA2-V` into `CAT045-CAT177-c713` in order to distinguish it from `CAT419-CAT177-c420`. The result is shown in Figure 7.10.

The remainder of Figure 7.6 attempts to give a sense of the kinds of constructions eventually learned by the learner. One of the last ten generalizations to be made by the learner in this experiment, `CAT981-CAT419-c1834` is a fairly general construction that describes an event in which the addressee (also a child) performs some continuous force application action (e.g. bring, carry, grasp) on any manipulable object.

Quantitative results

It is certainly the task of a grammar learner to learn a set of reasonable constructions, but its most important goal of all is to get better at interpreting utterances and to rely less on the context during the interpretation process. Another obvious goal is for the learner to be able to produce more communicative sentences, but without the support of a model of production, the evaluation strategies of the current work has to be based on the model's comprehension ability on seen and unseen data. Unfortunately there are no well-established benchmarks for evaluating grammar learning on a child language corpus in a construction grammar framework, so a few numerical measures are used here to assure the reader that the model is working fairly well. The first is a sanity check using a scrambled corpus, the second is a measure of semantic accuracy using the gold standard semantic annotation, and the third is a measure of analysis cohesiveness using the average number of roots per analysis.

Test 1: Preference procedure with scrambled corpus

As is consistent with preference procedures in behaviorial language learning experiments, we believe that a learner ought to prefer grammatical sentences to ungrammatical ones if it has successfully learned something about the structure of a language. Although this simplistic test does not measure the learner's comprehension ability, any successful learner should pass this test. This is carried out in the model by scrambling both validation sets so that words in each utterance appear out of their normal order. The best-fit analyzer will then use the initial grammar as well as the final grammar to analyze both the original and the scrambled corpus. Recall from Section 2.3 that the analyzer tries to find the most likely interpretation given the grammar, an utterance and its surrounding context, i.e. $a = \underset{a}{\operatorname{argmax}} \mathrm{P}(a \mid sentence, grammar, context)$.

During the process, it assigns each utterance a score based on its log probability; the total score of the corpus is the sum of the log probabilities. Given the initial grammar (i.e. only the lexicon), we expect the learner to make little or no distinctions between the original and the scrambled corpus, whereas after successful learning, we expect the log probability of the original corpus to be considerably higher than that of the scrambled corpus.

Figure 7.13 and Figure 7.14 show the total log probability of the training corpus and the unseen validation corpus before and after learning. One utterance in each corpus had to be thrown out because the analyzer garden-pathed and could not recover before running out of memory. In all cases the total log probability of the data increased with learning: this is due to the fact that the analyzer is able to parse robustly and extract bigger fragments of analysis even if the data is scrambled. Consider a scrambled sentence such as *Xi1xi1 yao4 chi1* (Xixi medicine eat). An analyzer that has a construction such as `XI1XI-CHI1-c028` is able to recognize *Xi1xi1 chi1* with one construction while skipping over *yao4*. Both the original and scrambled utterances can be analyzed with two roots. Additionally, compared to a lexicon-only grammar, the analysis of the scrambled utterance still incurs less multi-root penalty (see Section 2.3) and has a higher probability for each of the fragment of analysis.

As predicted, the increase in log probability is greater in magnitude in the seen data than the unseen data. Crucially, the learning does lead the model to prefer the original corpus to the scrambled corpus, as evident both by the total log probability of each corpus after learning as well as the average increase in log probability per utterance before and after learning.

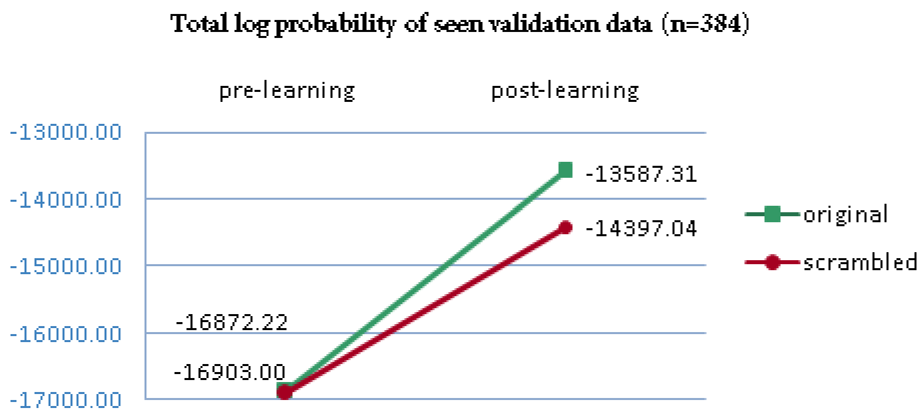

Figure 7.11 The total log probability assigned to the original and scrambled seen validation data.

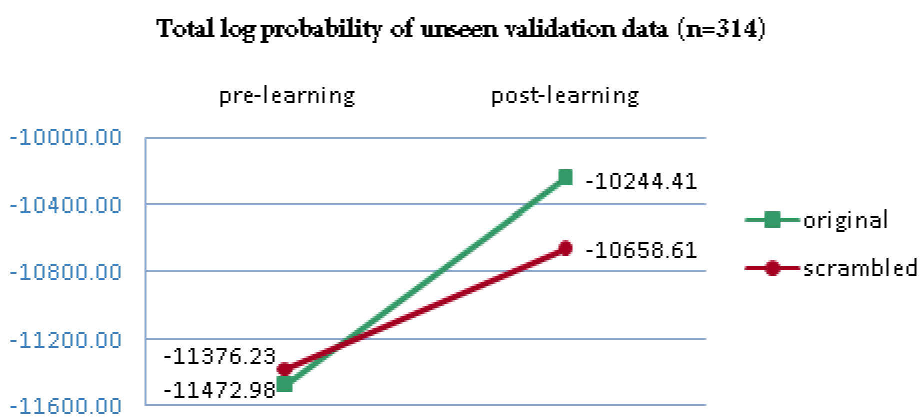

Figure 7.12 The total log probability assigned to the original and scrambled unseen validation data.

The average increase in log probability per utterance in the seen validation data is 8.55 for the original corpus and 6.53 for the scrambled corpus. The average increase in log probability per utterance in the unseen validation data is 3.91 for the original corpus and 2.29 for the scrambled corpus. Out of the 384 utterances in the original seen validation data, 76 of them are recognized as more likely than their scrambled counterpart, 10 are recognized as less likely, and the rest make no difference. Out of the 314 utterances in the original unseen validation data, 34 of them are recognized as more likely than their scrambled counterpart, 10 are recognized as less likely, and the rest make no difference. Again, though not a definitive measure of the quality of the learned grammar by itself, this result suggests that the model is going in the right direction. We will next look at the gold standard scoring of the analyses of the two validation sets.

Test 2: Gold standard scoring of returned analyses

As described in Section 7.2, both the seen and unseen validation sets have gold standard annotations which represent the interpretation that the analyzer, using the grammar, minimally needs to extract from the utterances. Each annotation contains constituent bracket and semantic filler information for core verb arguments and core argument structure arguments.

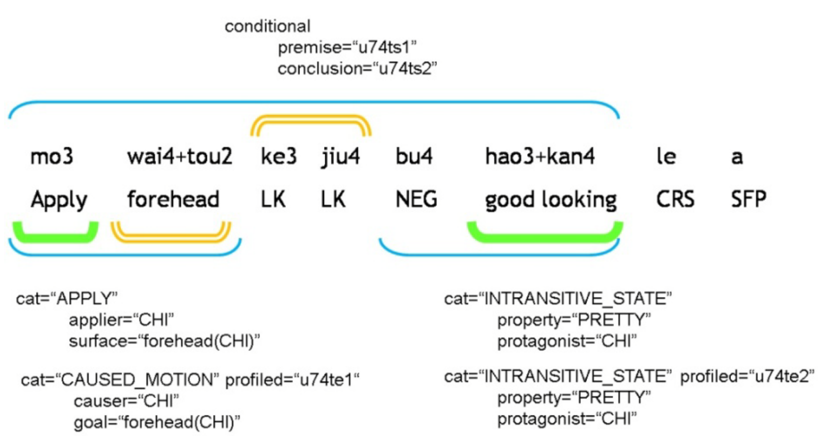

Figure 7.13 (reproduced from Chapter 2) Gold standard annotation of the utterance *mo3 wai4+tou2 ke3 jiu4 bu4 hao3+kan4 le a* (if you apply [the lotion] to your forehead then you won't be pretty). Both verb arguments and argument structure (phrasal) arguments are annotated, as shown in the bottom four annotations for the two clauses. Bracketing information is supplied (verb brackets shown in bolded lines, phrasal brackets in thin lines, and any additional words or arguments in double lines) as well as any interesting sentential constructions (e.g. conditionals).

Combining verb arguments and argument structure arguments in the same scoring mechanism, the gold standard annotation naturally affords 4 ways to score a particular analysis. These four scores and an example of each are:

- syntactic bracketing score

  Does the analysis supply the `surface` *role of the* APPLY *schema with a word at (1, 2)?*

- event core argument semantic type score

  Is the type restriction of the `surface` *role of the* APPLY *schema in the semspec compatible with* `@Forehead` *(e.g.* `@Body_Part` or `@Physical_Object`*)* [36]*?*

- event core argument resolution score

  Is the child's forehead proposed by the analyzer as a possible referent of the `surface` *role?*

- event core argument contextual fit score

  Is the child's forehead exactly the entity chosen by the context-fitter to be the `surface`*?*

Due to idiosyncrasies of the scoring method (as explained in the footnote), the core argument semantic type score is not a measure sensitive to the progress of early grammar learning. Similarly, the ability of the context fitter to tie the utterance to context, as reflected by the

---

[36] Notice that the linguistically-supplied semantic type is almost always less specific than the entity found in context (i.e. the annotated entity) due to say, pronoun use. The best that the scorer can do is to check that the slot type constraint found in the semspec is a supertype of the gold-standard type. This is a big obstacle in determining the semantic correctness of the analysis and renders this approach entirely unsuitable for evaluating grammar learning progress. Here's why:

The type constraints given by the schemas on the roles are the most general and are therefore always supertypes of the annotated types (e.g. the APPLY schema only requires the surface to be a `@Physical_Object`). With only lexical constructions in the grammar we expect the verb core argument score to be fairly close to 1 (and in practice lower due to lexical ambiguity), which is counter-intuitive. As the grammar gets more complicated, different roles are unified, contextual constraints are learned, and the type constraints in the corresponding slots get stricter and less likely to be correct. This score is therefore expected to decrease initially as mistakes are made and slowly come back up to close to 1 as the grammar becomes more adult-like. With the focus of the current model being early constructions, the semantic argument score is inappropriate. It does, however, work for the analyzer task in Chapter 2 because comparable adult-like grammars are used.

The argument structure core argument score is expected to be at 0 with a lexicon-only grammar, but there are so few of these argument structure core arguments that do not overlap with the verb arguments that this measure is unrepresentative as well.

contextual fit score, is also largely independent of the early fluctuations in grammar learning. This leaves two scoring criteria for the experiment — the syntactic bracketing score and the resolution score, combined between verb argument and argument structure arguments — both of which are reported as f-scores, i.e. the harmonic mean of precision and recall. Since this is not a standard information retrieval task in that the semantic analysis is expected to yield more relations than are given in the gold standard, a modified definition of precision and recall is used:

$$precision = \frac{\#correct}{\#\ correct + \#\ incorrect}$$

$$recall = \frac{\#correct}{\#\ correct + \#\ incorrect + \#\ nomatch}$$

where *#correct* is the number of answers (brackets or resolution) that the analysis gets correct out of those expected by the gold standard, *#incorrect* is the number of answers that the analysis gets incorrect out of those expected by the gold standard, and *# nomatch* is the number of answers that the gold standard expects but are not present in the analysis. The f-score is

$$\frac{2 \cdot precision\ \cdot recall}{precision\ +\ recall}$$

as is standard. Both the syntactic bracketing score and the resolution score are somewhat generous but they provide a reasonable proxy for determining whether the model is able to extract increasingly richer interpretation of the utterances based on the learned grammars.

The constituent bracketing and event core argument resolution scores attained by the model are reported in Figure 7.15 and Figure 7.16. After only 750 learning episodes, the model improves on both measures on both the seen and unseen validation sets. For the constituent bracketing score, precision goes down after learning not unexpectedly: the model initially gets (most) main verb brackets correct just based on the lexical constructions and misses everything else. After learning, the model is able to suggest a good number of core argument brackets as well, some of them incorrectly. For the core argument resolution, both precision and recall go up. Due to the fact that the training was cut short, the coverage of the grammar is still fairly limited as evident in the low recall. The progress of learning can be seen in Figure 7.17 which reports the resolution scores that are recorded every 50 episodes.

	constituent bracketing			event core argument resolution		
	precision	recall	f-score	precision	recall	f-score
pre-learning	1.000	0.454	0.625	0.000	0.000	0.000
post-learning	0.973	0.547	0.700	0.661	0.243	0.355

Figure 7.14 Seen validation data: constituent bracketing scores and core argument resolution scores.

	constituent bracketing			event core argument resolution		
	precision	recall	f-score	precision	recall	f-score
pre-learning	1.000	0.424	0.595	0.000	0.000	0.000
post-learning	0.989	0.505	0.669	0.492	0.121	0.194

Figure 7.15 Unseen validation data: constituent bracketing scores and core argument resolution scores.

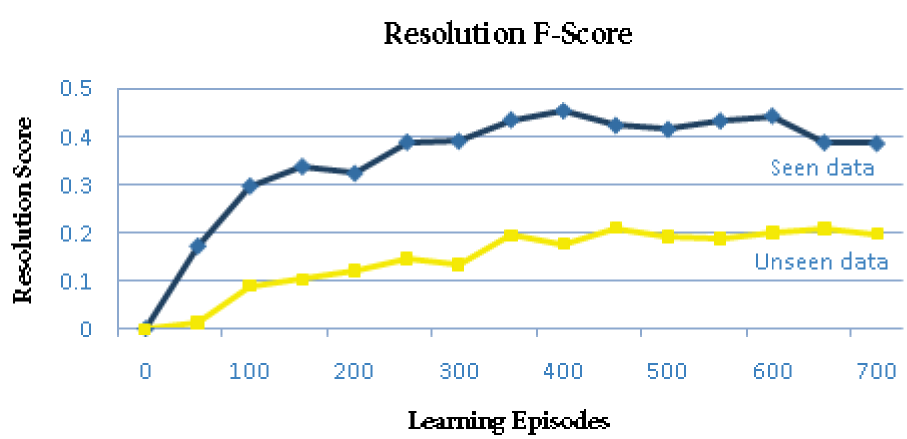

Figure 7.16 The resolution score on the seen and unseen validation sets as recorded throughout the learning experiment.

Test 3: Multi-rootedness of returned analyses

One final measure used here is the average number of roots per analysis, which is the number of top-level constructions needed to cover the entire utterance. For example, a total of four constructions are used in both analysis shown in Figure 7.12. However, the analysis on the left has one root, XI1XI1-CHI1-YAO4-c040 whereas the one on the right has two roots, XI1XI-CHI1-c028 and YAO4-N. We expect that prior to learning, the average number of roots in the analyses is equal to the mean length of utterance in the corpus, and will slowly decrease as the grammar gets more sophisticated. With an adult grammar we expect the average number of roots to approach 1.

Figure 7.17 On the left, an analysis with one root. On the right, an analysis with two roots.

Tthe average number of roots per analysis in both validation sets is shown in Figure 7.18. At the beginning of the learning experiments, the average number of roots is close to the parental MLU [37] and gradually decreases as learning progresses.

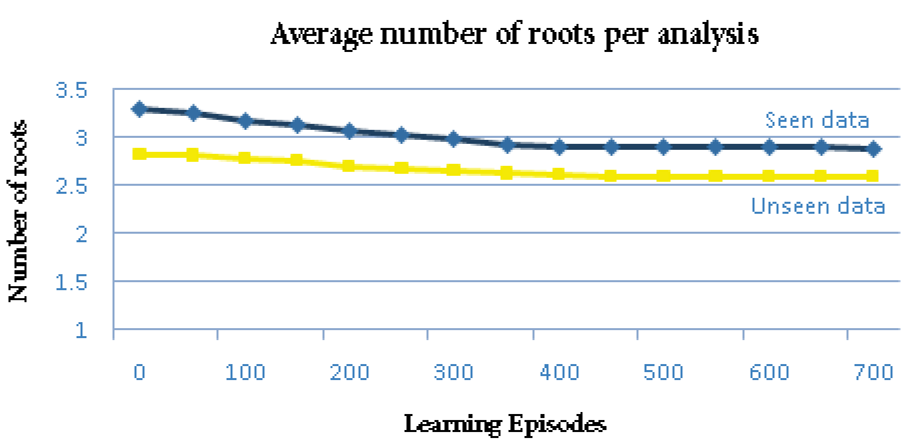

Figure 7.18 The average number of roots per analysis in the seen and unseen validation set throughout the learning experiment.

[37] It is a little lower because of a construction in the initial grammar that allows the analyzer to chunk consecutive interjections.

7.3 Experiment 2: Model variations on the Mandarin CHILDES data

As is apparent from Experiment 1, the learned grammar quickly outgrew the capacity of the analyzer. This section describes a series of model manipulations to examine the behavior of the model at the macro level using quantitative measures such as grammar size and resolution score as a funtion of learning episodes. The first manipulation attempts to control the size of the grammar by enabling the decay operation. The second manipulation examines the effect that the grammar statistics update has on the model's ability to generalize to unseen data by changing the statistics update discount factorγ to 0.2. The third manipulation looks at the contribution of a good context-fitting mechanism by way of utilizing the gold standard annotation in learning.

Variation 1: enabling decay

In this first variation, decay was enabled and set to purge any construction which is last modified over 50 learning episodes ago and which has been used fewer than 3 times total. The rest of the model remains unchanged and the same training procedure as the basic model was used. Figure 7.19 shows the growth in size of the grammars in the "with-decay" model as compared to the basic model. After 750 learning episodes, the basic model had 490 learned concrete constructions and 18 learned abstract categories in its grammar. In contrast the grammar in the "with-decay" model had 176 concrete constructions and 14 abstract categories at the same point in time. The grammar in the "with decay" model continued to grow with training input, reaching close to 410 concrete constructions by the time the model ran out of memory and had to be aborted at episode 1700.

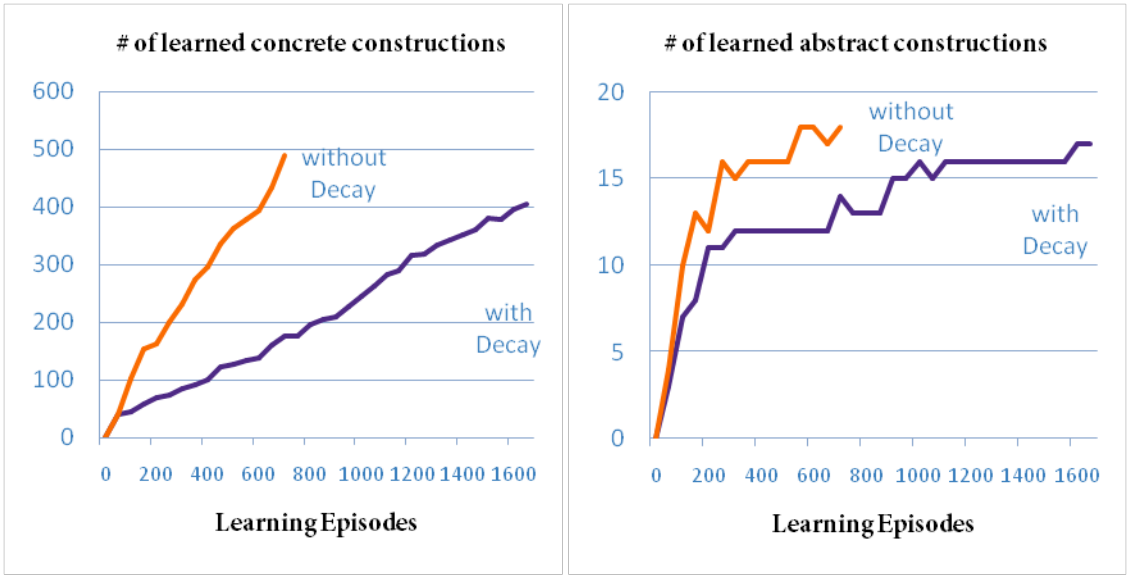

Figure 7.19 The size of the grammar grows more slowly with the decay operation enabled in the model. The graph reflects the number of learned concrete or abstract constructions that are retained in the grammars at any given point (and not the total that it has ever tried to learn).

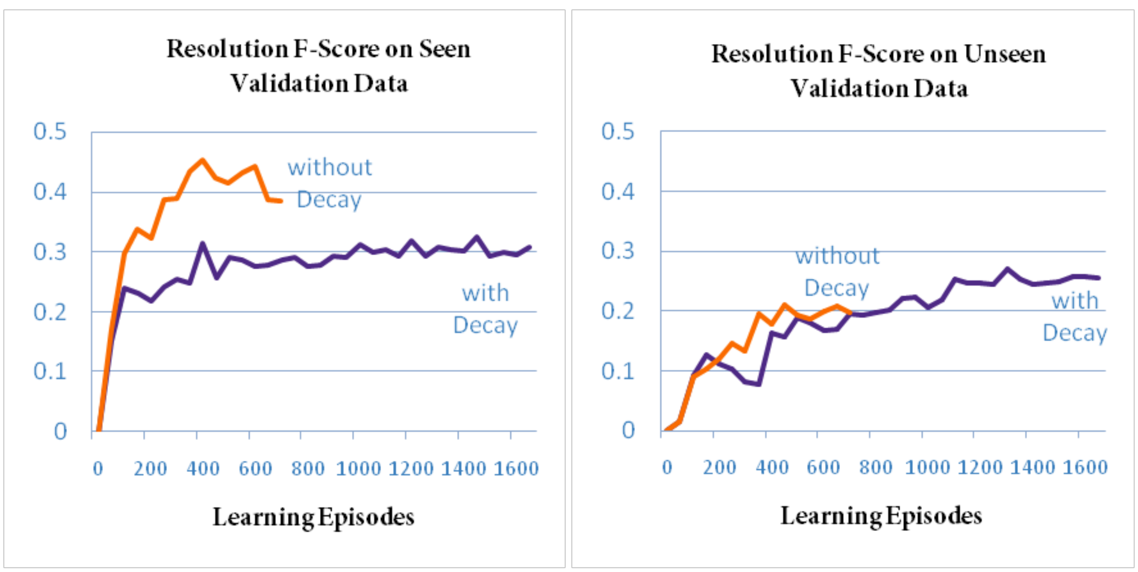

Figure 7.20 Resolution scores attained by the learned grammars on the seen and unseen validation data, with and without the decay operation enabled. The score on the unseen data is largely unaffected by the decay operation and continues to improve long after the basic model has to be aborted.

What is interesting here is the performance of the two models on the seen and unseen validation data. Figure 7.20 shows the resolution score of the "with-decay" model in comparison 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.

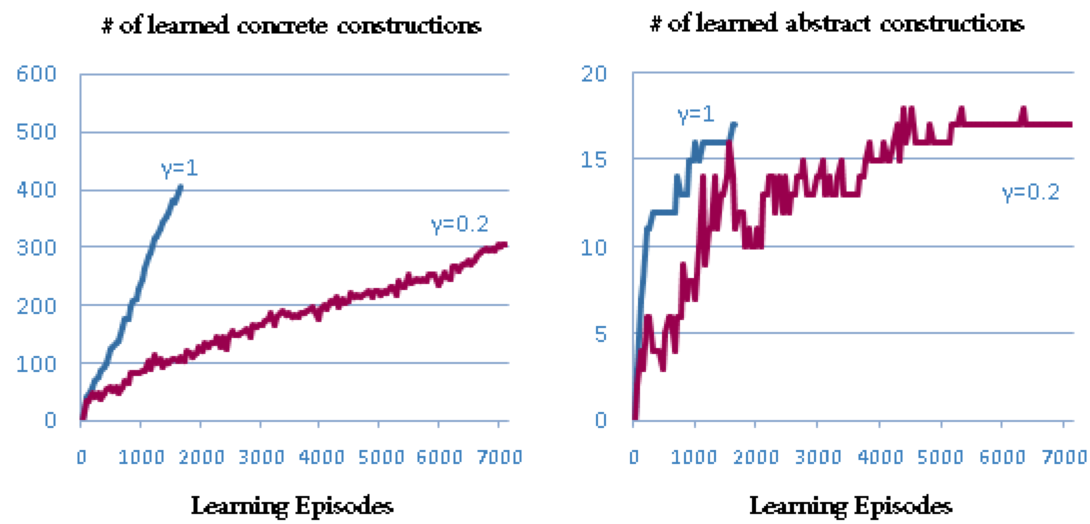

Figure 7.21 The size of the grammar grows with more fluctuations when γ = 0.2 than when γ = 1.0. The grammar seems to stabilize right at around 17 abstract categories though the number of concrete constructions continues to increase throughout.

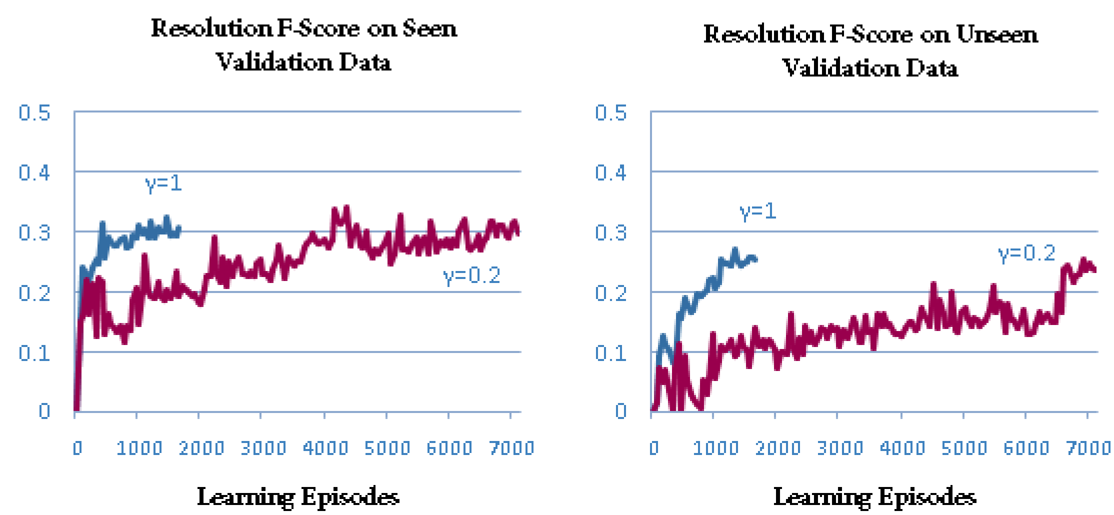

Figure 7.22 The resolution scores obtained on both datasets with γ = 0.2 are comparable to those with γ = 1.0 even though it takes the learner much longer to arrive at that level of performance.

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 "γ = 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 "γ = 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 "γ = 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 "γ = 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 "γ = 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

$\gamma = 0.2$	$\gamma = 1.0$
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 γ = 0.2 model are shown on the left, and the 17 categories in the grammar learned by the γ = 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.

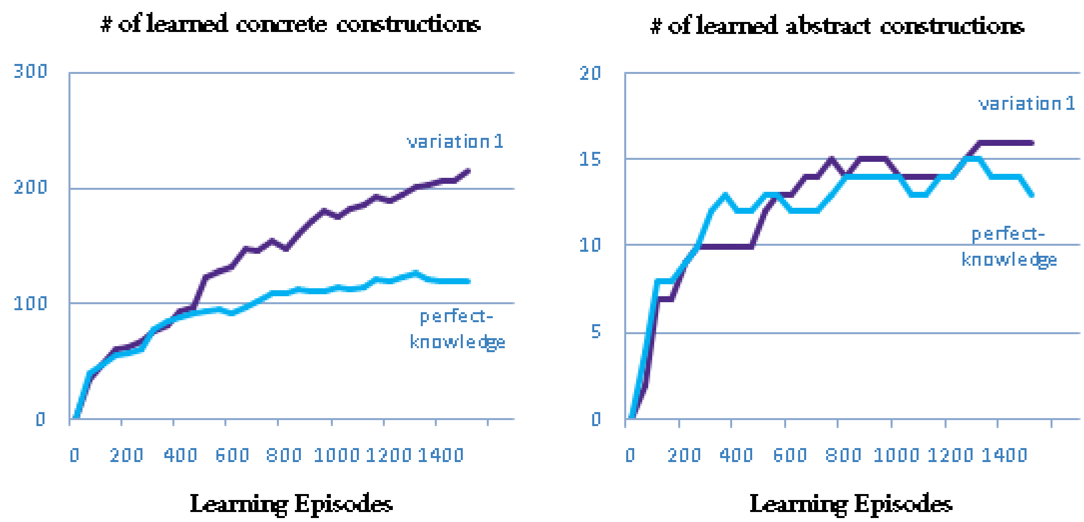

Figure 7.25 The number of learned concrete constructions after 4 learning iterations is significantly higher for the variation 1 model than the "perfect-knowledge" model, reflecting the amount of spurious constructions in the grammar.

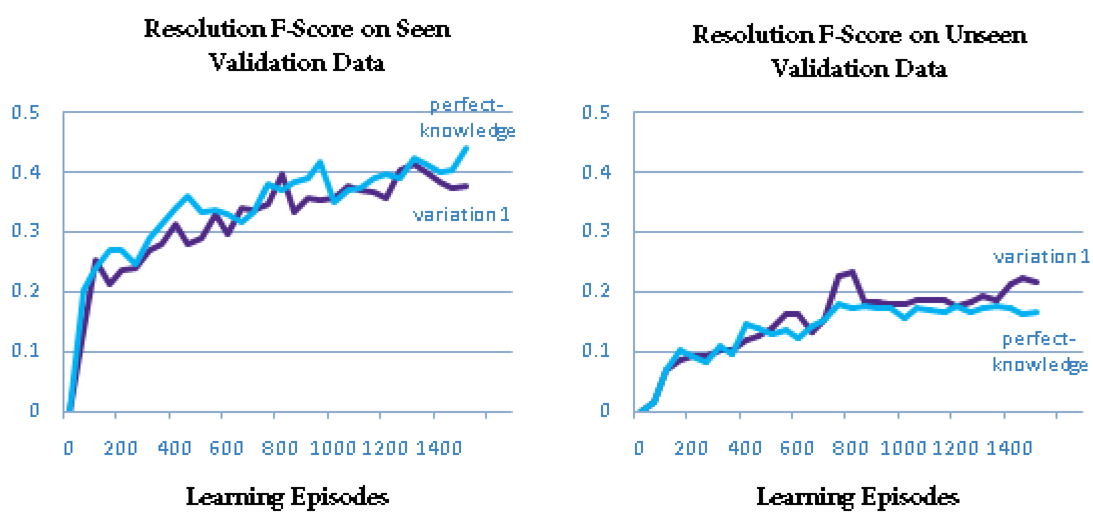

Figure 7.26 The resolution score on the seen and unseen validation data by the variation 1 vs "perfect-knowledge" model.

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	Haoyu disturb horse	find fish
sleep	throw	give car
like rice	Aunt throw	I give
I like fish	take car	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.

iter	**I** **Comp**	II Comp+Gen	III CG+Dec	IV CGD +Exp	V CGD+Rev	**VI** **CGD +Oms**	VII 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.

iter	I **Comp**	II Comp+Gen	III CG+Dec	IV **CGD +Exp**	V CGD+Rev	VI CGD +Oms	VII 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.

iter	**I** **Comp**	II Comp+Gen	**III** **CG+Dec**	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	VII 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.

iter	**I** **Comp**	**II** **Comp+Gen**	III CG+Dec	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	VII 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.

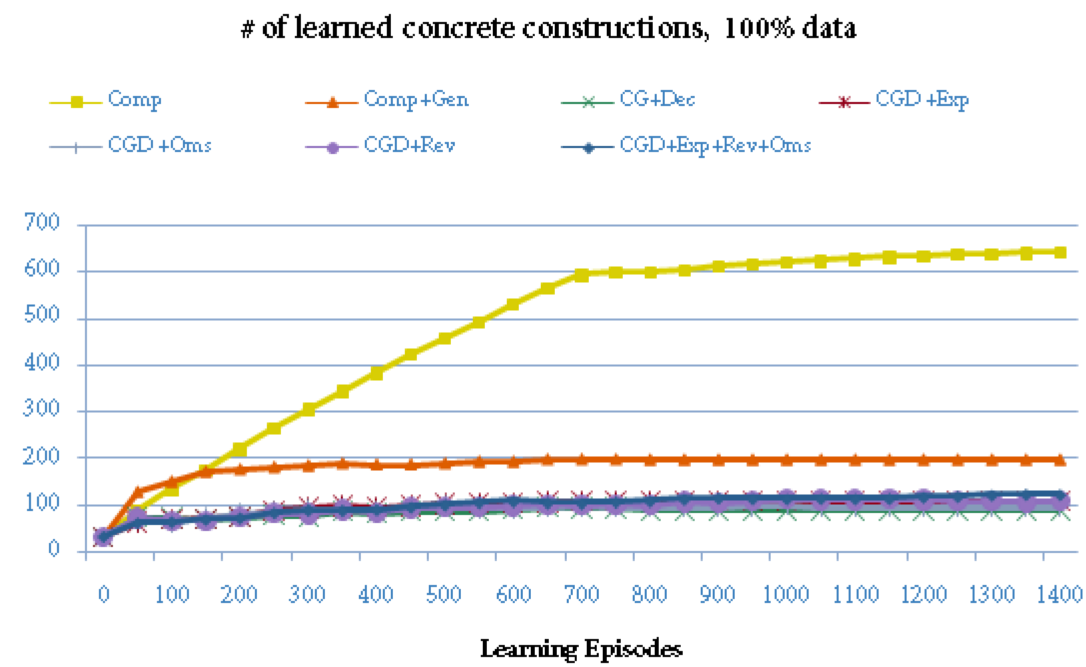

Figure 8.8 The number of learned concrete constructions as a function of the learning episodes.

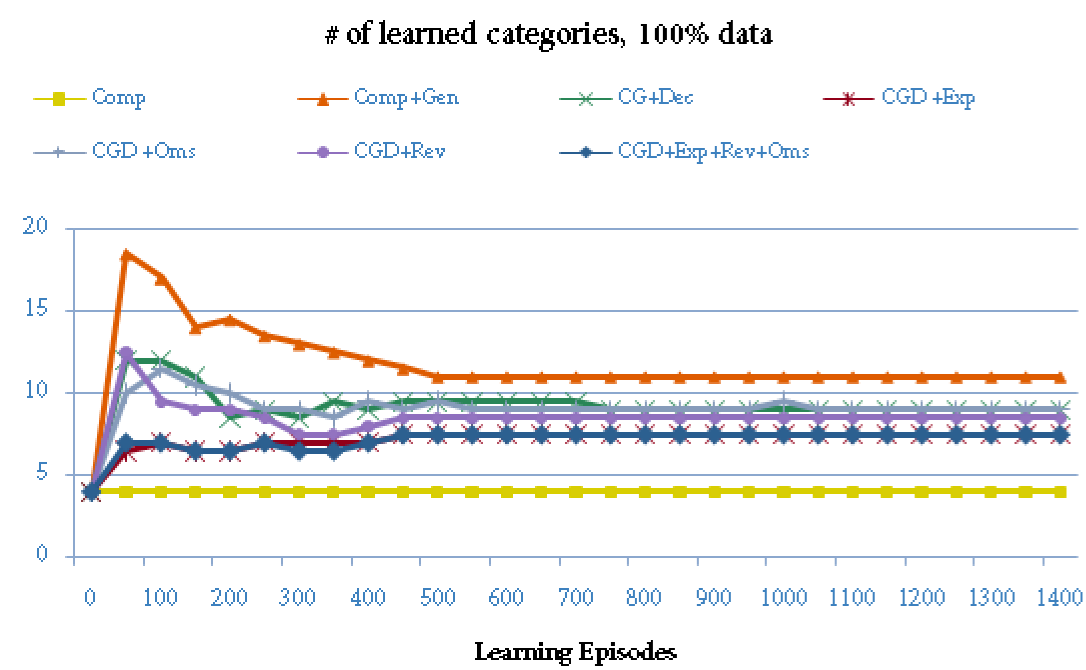

Figure 8.9 The number of learned categories as a function of learning episodes.

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, $P(expressed) = 0.3$ and $P(local \mid expressed) = 1.0$. For the recipient constituent, $P(expressed) = 0.4$ and $P(local \mid expressed) = 1.0$. 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, $P(expressed) = 0.4$ and $P(local \mid expressed) = 0.65$. 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 $P(expressedtheme) = 0.6$, and if the fronting conditions are met (i.e. the recipient is also expressed), $P(localtheme \mid expressedtheme) = 0.35$. 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 car throw	ba3 bear give you
ba3 Dad disturb	ba3 veggies find	Haoyu ba3 rice give I
HaoYu ba3 I disturb	you ba3 stool find	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.

iter	**I** **Comp**	II Comp+Gen	III CG+Dec	IV CGD +Exp	V CGD+Rev	**VI** **CGD +Oms**	VII 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.

iter	**I** **Comp**	II Comp+Gen	III CG+Dec	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	**VII** **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.

% data	IV CGD +Exp	V CGD+Rev	VI CGD +Oms	VII CGD+ Exp+Rev+Oms	Avg	Exp 3 Avg
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%.

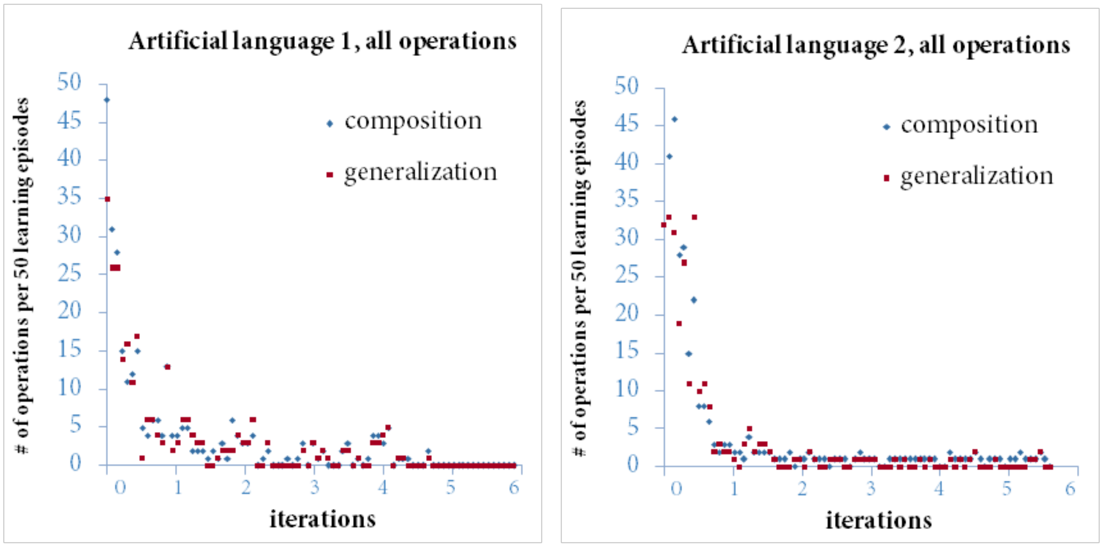

Figure 9.1 The number of composition and generalization operations performed by the learner decreases over time as the learner's grammar gets better and better at analyzing the learning input. Both graphs show the number of composition and generalization operations performed per 50 episode intervals 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.

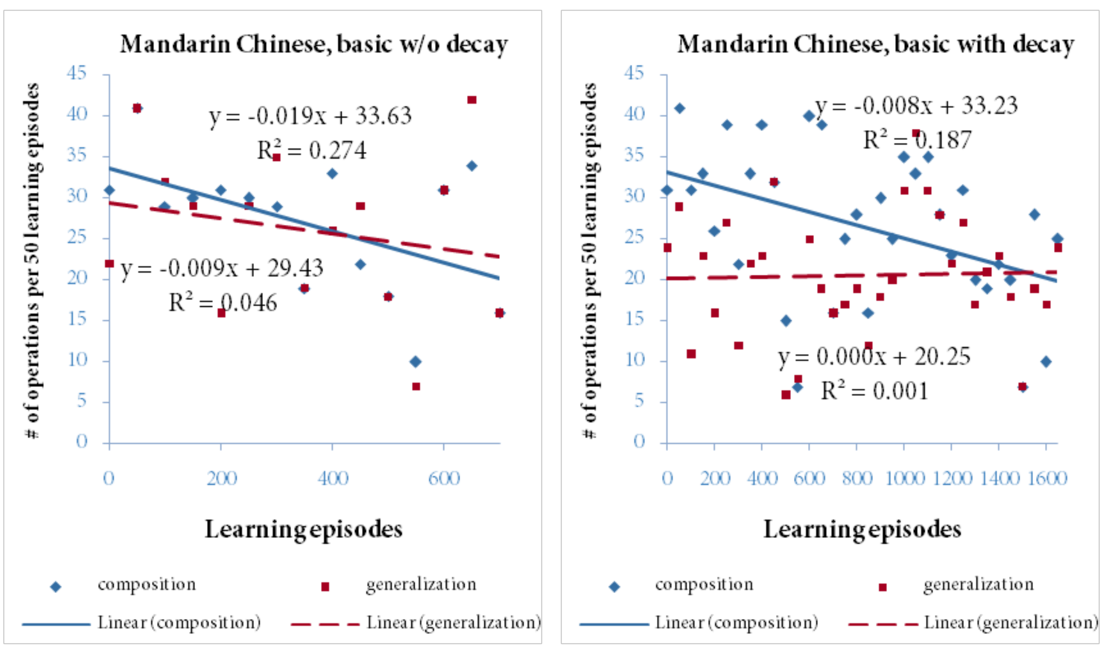

Figure 9.2 Both graphs show the number of composition and generalization operations taken per 50 episode intervals within the first iteration on the Mandarin Chinese corpus. The left shows data from the basic model without decay; the right shows data from the variation 1 model with decay. There is a 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 γ 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 α, loosely defined as the syntactic and semantic configuration of each construction that takes α as a constituent. For example category α 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 α is a "proto-subject", so to speak, and we will call this configuration of form and meaning relations its constructional context. Another category β may be another category of human nouns and may even share a number of members with α, but if β 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 α'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 α 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 α 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 α will be used as a constituent in the new GEI3-α construction. The resulting category α 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:

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

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.

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

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	$CV_{instrument}$ - 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 - $CV_{benefactive}$	
XIA4-DI4-c396	$LOC_{downwards}$ - ground	
HUI4-CHI1-c616	able - eat	yes
LIANG3-GE-c625	two - CLS	yes
QI3-LAI2-c757	rise - DIR_{up}	yes
WANR2-NE-SP-c787	play - SFP	?
NEI3-GE-c824	that - CLS	yes
GUAI3-GUO4-c849	turn - DIR_{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))

<table>
<tr><td>ECGL</td><td>→</td><td>ECGL Schema | ECGL Cxn | ε</td></tr>
<tr><td>CxnKind</td><td>→</td><td>GENERAL CONSTRUCTION | CONSTRUCTION</td></tr>
<tr><td>Cxn</td><td>→</td><td>CxnKind IDENT ParentLOpt CBlockOpt FBlockOpt MBlockOpt</td></tr>
<tr><td>BlockType</td><td>→</td><td>: Typespec | ε</td></tr>
<tr><td>CBlockOpt</td><td>→</td><td>CONSTRUCTIONAL BlockType ConstitsLOpt ConstraintLOpt | ε</td></tr>
<tr><td>FBlockOpt</td><td>→</td><td>FORM BlockType ConstraintLOpt | ε</td></tr>
<tr><td>MBlockOpt</td><td>→</td><td>MEANING BlockType EvokedLOpt RolesLOpt ConstraintLOpt | ε</td></tr>
<tr><td>SchemaKinds</td><td>→</td><td>FEATURE SCHEMA | SEMANTIC SCHEMA | SCHEMA</td></tr>
<tr><td>Schema</td><td>→</td><td>SchemaKinds IDENT ParentLOpt EvokedLOpt ...
RolesLOpt ConstraintLOpt</td></tr>
<tr><td>ParentLOpt</td><td>→</td><td>ParentL | ε</td></tr>
<tr><td>SubcaseOf</td><td>→</td><td>SUBCASE | SUBCASE OF</td></tr>
<tr><td>ParentL</td><td>→</td><td>ParentL , IDENT | SubcaseOf IDENT</td></tr>
<tr><td>Typespec</td><td>→</td><td>IDENT | EXTERNALTYPE</td></tr>
<tr><td>EvokedElement</td><td>→</td><td>EVOKES Typespec AS IDENT</td></tr>
<tr><td>EvokedLOpt</td><td>→</td><td>EvokedLOpt EvokedElement | ε</td></tr>
<tr><td>Role</td><td>→</td><td>IDENT OptType</td></tr>
<tr><td>OptType</td><td>→</td><td>: Typespec | ε</td></tr>
<tr><td>RolesLOpt</td><td>→</td><td>RolesL | ε</td></tr>
<tr><td>RolesL</td><td>→</td><td>RolesL Role | ROLES</td></tr>
<tr><td>Constit</td><td>→</td><td>OPTIONAL IDENT : IDENT OptConstitAnno</td></tr>
<tr><td></td><td>|</td><td>EXTRAPOSED IDENT : IDENT OptConstitAnno</td></tr>
<tr><td></td><td>|</td><td>IDENT : IDENT OptConstitAnno</td></tr>
<tr><td>OptConstitAnno</td><td>→</td><td>[ProbL] | ε</td></tr>
<tr><td>ProbL</td><td>→</td><td>PROB | PROB , PROB</td></tr>
<tr><td>ConstitsLOpt</td><td>→</td><td>ConstitsL | ε</td></tr>
<tr><td>ConstitsL</td><td>→</td><td>ConstitsL Constit | CONSTITUENTS</td></tr>
<tr><td>ConstraintLOpt</td><td>→</td><td>ConstraintL | ε</td></tr>
<tr><td>ConstraintL</td><td>→</td><td>ConstraintL OptIgnore Constraint | CONSTRAINTS</td></tr>
<tr><td>ChainOperator</td><td>→</td><td>⟷ | BEFORE | MEETS</td></tr>
<tr><td>OptIgnore</td><td>→</td><td>IGNORE | ε</td></tr>
<tr><td>Var</td><td>→</td><td>SLOTCHAIN | IDENT</td></tr>
<tr><td>Constraint</td><td>→</td><td>Var ChainOperator Var | Var ⟵ IdentOrStr</td></tr>
<tr><td>IdentOrStr</td><td>→</td><td>EXTERNALTYPE | IDENT | STR</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>IDENT</td><td>→</td><td>[A-Za-z][0-9a-zA-Z-_]*</td></tr>
<tr><td>SLOTCHAIN</td><td>→</td><td>(IDENT.)+IDENT</td></tr>
<tr><td>EXTERNALTYPE</td><td>→</td><td>@[0-9a-zA-Z.-_]+</td></tr>
<tr><td>STR</td><td>→</td><td>“(\” | [∧\n”] | \{WHITE_SPACE_CHAR}+\)*”</td></tr>
<tr><td>PROB</td><td>→</td><td>.[0-9]+ | 1.0 | 1</td></tr>
</table>

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>
```