Chapter 5

Learning embodied constructions

It isn't that they can't see the solution. It's that they can't see the problem.
— G.K. Chesterton

We are now in a position to address ourselves more directly to the learning task introduced in Chapter 1. Section 5.1 recapitulates the most relevant constraints encountered in the foregoing chapters. These are distilled in Section 5.2 into a formal statement of a class of grammar learning problems, along with a specific instance of this class. The focus then shifts to seeking an adequate solution: Section 5.3 reviews some candidate approaches, and Section 5.4 lays out a framework that adapts these to satisfy our problem constraints.

5.1 The child's learning task: a review

The developmental and linguistic evidence reviewed thus far suggests the following informal synopsis of the language learning task: Children are situated in rich experiential contexts, subject to the flow of unfolding events. At all stages, they exploit a panoply of cognitive and social abilities to make sense of these experiences. To make sense of *linguistic* events — sounds and gestures used in their environments for communicative purposes — they also draw on mappings between these perceived forms and conceptual knowledge. These mappings, or **constructions**, are typically incomplete with respect to the utterances they encounter, but even an imperfect understanding of an utterance provides direct clues to its intended meaning and thereby reduces the burden on pragmatic inference and situational context. The goal of language learning, from this perspective, is to acquire an increasingly *useful* set of constructions (or **grammar**) — that is, one that allows accurate interpretations of utterances in context with minimal recourse to general inference procedures. In the limit, the grammar should stabilize, while facilitating the comprehension of both familiar and novel input.

This section summarizes the most relevant constraints on the problem. These fall into three categories: structural constraints arising from the nature of the target of learning; usage-based constraints arising from the goal of facilitating language comprehension; and cognitive and developmental constraints arising from the nature of the learner.

5.1.1 Representational requirements

The target of learning is a construction-based grammar in the general sense described in Section 2.2.1, as instantiated by the Embodied Construction Grammar formalism.

- Constructions are *cross-domain*: they are mappings over the domains of form and meaning. Units of form and meaning are *schematic* descriptions summarizing the linguistically relevant aspects of more detailed embodied representations (such as acoustic or auditory representations for form, and motor, perceptual and other conceptual representations for meaning).
- Constructions may have *constituents*, *i.e.*, subcomponents that are themselves constructions.
- Constructions are themselves complex *categories*, organized in a *typed multiple inheritance hierarchy*. They may be subcases of other constructions, inheriting structures and constraints. Their constituent constructions, as well as their form and meaning poles, may also be typed.[1]
- Constructions may include *relational constraints*, i.e. relations that must hold of form and/or meaning elements (such as binding constraints and ordering constraints).

The focus here is on constructions licensing multiword expressions, which typically involve multiple forms, meanings and constituents, along with relational constraints among these. The problem thus demands an approach to learning that can accommodate the representational complexity imposed by structured cross-domain mappings.

[1] As noted in Section 3.1.4, the choice of an inheritance-based type hierarchy, while a reasonable simplification for current purposes, does not capture the graded, radial nature of constructional organization.

5.1.2 Comprehension-based constraints

The framework described in Chapter 4 distinguishes several processes involved in language comprehension:

- *constructional analysis*: identifying the set of constructions instantiated by a given utterance, and its corresponding *semantic specification* (or *semspec*);
- *contextual resolution*: mapping objects and events in the semspec to items in the current communicative context, producing a *resolution map* and a *resolved semspec*; and
- *simulative inference*: invoking the dynamic embodied structures specified by the (resolved) semspec to yield contextually appropriate inferences.

As discussed in Chapter 4, these processes must tolerate uncertainty, ambiguity and noise: there may be multiple possible constructional analyses for a given utterance and multiple ways of resolving referents to the context; there may be errors in the perceived input utterance or communicative context; and some inferences may be only probabilistically licensed. Comprehension processes must also cope gracefully with input not covered by its current grammar: all constructional analyses, including those that account for the entire input utterance and those that do not (*i.e.*, *partial* analyses), should produce (partial) semspecs that can be resolved and simulated.

These considerations suggest that comprehension is not a binary matter, but rather one of degree: interpretations can be judged as relatively more or less complete, coherent and consistent with the context. That is, the language comprehension processes above require some means of evaluating candidate interpretations and choosing those that contribute to effective and efficient comprehension (*e.g.*, by maximizing utterance interpretability in context, or minimizing constructional and contextual ambiguity). Since progress in learning can be judged only by improvement in comprehension ability, the same facilitating factors and evaluation criteria should also serve to guide and constrain the learning process.

5.1.3 Developmental desiderata

As reviewed in Section 2.1.3, children entering the multiword stage have amassed a battery of sensorimotor, conceptual, pragmatic and communicative skills, including:

- familiarity with a variety of people, objects, locations, and actions, including both specific known entities (*e.g.*, mother, favorite stuffed toy, bed) and more general categories of objects and actions (*e.g.*, milk, blocks, grasping, being picked up);

- ability to infer referential and communicative intent, including determining objects of (joint) attention and speech act type;
- familiarity with intonational contours and their associated speech acts;
- relatively well-developed ability to segment utterances into words; and
- a growing inventory of lexical mappings, including labels for many familiar ontological categories, as well as social and relational words.

The timelines for these ongoing developmental and learning processes overlap with the acquisition of multiword constructions, and they vary dramatically across individual children. But all of these diverse kinds of information may, in principle, be available by the time children begin to learn their first word combinations.

Several trends in the acquisition of multiword constructions have also been identified, as reviewed in Section 2.2.3. Typically, children:

- learn more specific constructions before more general constructions, more frequent constructions before less frequent constructions, and smaller constructions before larger constructions;
- require a relatively modest amount of data to achieve a baseline level of performance;
- receive relatively little negative evidence in the form of error correction; and
- generalize constructions to arguments beyond those observed in the input, both appropriately and inappropriately.

Again, these patterns are subject to variation, both across and within individuals. Some children learn larger multiword chunks as a single unit before later reanalyzing them in terms of their constituent pieces; some persist in producing one-word utterances despite being able to string several of these together (separated prosodically) to express complex predications.

These developmental findings inform the kinds of behaviors that should be exhibited by a successful model of construction learning. That is, the course of acquisition should be qualitatively comparable to that of a child at a similar developmental stage, and the model should be flexible enough to encompass multiple learning styles.[2]

[2] I elide the distinction between comprehension and production here; while the timeline of acquisition may differ between these, I assume that many aspects of the course of acquisition are roughly comparable.

5.2 Problem statement(s)

This section translates the constraints just reviewed into a formal statement of a class of grammar learning problems. Section 5.2.1 identifies the main structures and processes involved (defined in detail in Section 4.1) and uses them to define the general problem of comprehension-based language learning; these are then elaborated with respect to the problem of learning relational constructions in Sections 5.2.2-5.2.5, culminating in a more specific problem statement in Section 5.2.6.

5.2.1 Construction learning defined

The primary representational structures involved in language learning are as follows:

- A **grammar** G is a pair (S, C), where S is a **schema set**, and C is a **construction set** defined with respect to S. Both S and C are typed multiple inheritance hierarchies whose members are represented using the ECG schema and construction formalisms, respectively.
- Schemas in S provide parameters to embodied structures in a separate **ontology** O of detailed sensorimotor representations and other structures that are not specifically linguistic.[3]
- A **corpus** D is an ordered list of **input token**s d. Each input token d is a pair (u, z), where u is an **utterance** and z is a communicative **context**.

Language comprehension is divided into several interrelated processes:

- A **construction analyzer** provides the function $\text{analyze}(d, G)$, which takes an input token $d = (u, z)$ and a grammar G and produces an **analysis** a_d and an associated **semspec** $ss_a = \text{semspec}(a_d)$.
- A **contextual resolver** provides the function $\text{resolve}(ss, z)$,which takes a semspec ss and a context z and produces a **resolution map** r and a **resolved semspec** rss.
- A **simulation engine** provides the function $\text{simulate}(ss, z)$, which takes a resolved semspec rss as input parameters for a dynamic simulation with respect to the context z, using structures in the conceptual ontology O to produce a set $\text{inferences}(rss)$ of simulation inferences.

These processes evaluate their output structures using the folllowing quantitative scoring criteria:[4]

[3] The sense of 'ontology' intended here is broader than that typically used to represent static taxonomic relations.

[4] Probabilistically minded readers may note a suggestive resemblance between the scoring functions listed here and the Bayesian prior probability of a grammar, likelihood of observed data with respect to a grammar, and posterior probability of a grammar given the data, respectively). This family resemblance is further elucidated in Chapter 7.

- $\text{score}(G)$ is the **grammar score**, which measures intrinsic properties of a grammar G. Generally, it is defined to reward simpler, more compact grammars.
- $\text{score}(d|G)$ is the **token score**, which measures how well token d is comprehended using grammar G. Generally, it is defined to reward simpler, more likely analyses that account for more of the input forms and meanings and are more easily interpreted in context.
- $\text{score}(D|G)$ is the **corpus score**, which aggregates token scores $\text{score}(d|G)$ based on grammar G over the entire corpus D.
- $\text{score}(G|D)$ is the **grammar performance score**, which measures how well a grammar G accounts for a corpus D. This term should incorporate aspects of both the (intrinsic) grammar score and its associated corpus score.

The terms above allow us to characterize a general class of language learning problems consistent with the goals and constraints established in the preceding chapters.

Comprehension-based language learning

hypothesis space The target of learning is a grammar G defined using the ECG formalism.

prior knowledge The learner has an initial grammar $G_0 = (S_0, C_0)$, a constructional analyzer that performs the analyze and resolve functions, a simulation engine that performs the simulate function, and a conceptual ontology O.

input data The learner encounters a sequence of input tokens from a training corpus D.

performance criteria The grammar performance score of successive grammars G_i should improve and eventually stabilize as more data is encountered.

The language learning problem as defined here admits many instantiations, corresponding to different ways in which the structures and processes involved can be further delineated. A version of the problem focusing on lexical learning, for example, may restrict the initial and target grammars to lexical constructions; the precise form of the input data (raw sensory data, phonological or phonemic sequences, orthographic strings) may be chosen to reflect differing assumptions about the perceptual abilities of the learner or the starting point of learning; and the scoring functions may vary in the aspects of the grammar and analysis they reward.

My focus in this work is on the comprehension-driven acquisition of relational constructions — *i.e.*, the emergence of constituent structure. I thus restrict each component of the broader language learning problem with some simplifying assumptions, as described in the next several sections.

5.2.2 Hypothesis space

The hypothesis space for language learning is defined as a grammar $G = (S, C)$, since both S and C are taken to be part of linguistic knowledge. Recall that the schema set S contains the basic units of linguistic form and meaning, as discussed in Section 3.2. These are the two domains linked by all linguistic constructions, which together encompass all *linguistically* relevant concepts, including both lexicalizable and grammaticizable notions in the sense discussed in Section 2.2.2 (Talmy 2000; Slobin 1997).

I distinguish such concepts from the set O of *all* ontologically available categories, concepts and domain-based knowledge. O is taken to encompass the full range of human conceptual and perceptual potential, from multimodal, biologically grounded structures used in action, perception, emotion and other aspects of experience to the more abstract but often richly structured categories and relations in less directly embodied domains. The schemas in S summarize (a subset of) the structures of O, and they exhibit noticeable regularities across languages (as might be expected based on their putative common embodied basis). In the current work, however, I assume that the actual set of linguistically relevant features, as well as the constellations of features that are regularly packaged for particular linguistic purposes, must be learned on a language-specific basis.

Different learning processes are associated with each of these structures:

- **concept learning**: the formation of *ontological categories* of O based on embodied experience;
- **(linguistic) schema learning**: the reification of ontological categories and features of O into the *linguistically relevant categories* of S, *i.e.*, a set of embodied schemas for the domains of form and meaning;
- **construction learning**: the pairing of form and meaning schemas of S into the *specifically linguistic categories* of C.

These processes occur in parallel through development and into adulthood, and the relationships among them are complex. There is a natural dependency of schema learning on concept learning (a schema being a specific kind of concept), and of construction learning on schema learning (a construction being a pairing of form and meaning schemas). But structures arising in each kind of learning may also exert a mutually reinforcing influence on those arising in the others. A strong Whorfian position, for example, might assert a determinative effect of S on O; a weaker stance might posit a more balanced relationship between them. Likewise, it may be precisely the regularity of certain form-meaning pairs in C that elevates their associated concepts into the schemas of S.

The current work does not attempt to model all of these concurrent interactions; rather, these processes are idealized as independent but interleaved. Specifically, I assume that construction learning can be usefully modeled as taking place with respect to a *fixed* schema set and conceptual ontology. The space of possible grammars G thus reduces to the space of possible construction sets C defined with respect to S, with particular emphasis on the acquisition of structured mappings.

Note that this formulation nonetheless presumes significant *indirect* interactions based on the inherent structural connections among concepts, schemas and constructions: constructions refer directly to form and meaning schemas, and indirectly (via schemas) to embodied structures and general ontological categories. Any changes to these structures (due to interleaved concept and schema learning) could potentially affect the course of construction learning, where shorter interleaving intervals reflect a tighter connection. In the limit, this interleaving could theoretically be on a per-token basis. Potential extensions to the model to permit more direct interactions among these processes are addressed in Chapter 9.

5.2.3 Prior knowledge

The learner is assumed to have access to both a grammar (with schemas and constructions) and a language understanding model (performing constructional analysis and contextual resolution); these correspond to naive theories of language structure and use appropriate for supporting language behaviors observed during the single-word stage. No claims are made here about whether or to what extent these structures and processes are innate, or simply acquired before and during single-word learning.

Schemas and constructions

Schemas and constructions are represented using the ECG formalism, as described in Chapter 3 and Section 4.1.1. Assumptions about the contents of the initial grammar depend on the particular phenomena or learning stage of interest; the specifications below reflect the current focus on the earliest relational constructions in English:

- Word forms are represented as orthographic strings. Intonational information is represented using an intonation role filled by one of a limited set of familiar contours (falling, rising and neutral). Potential form relations are limited to the ordering relations before and meets.

- Meaning schemas reflect a typical child's ontological knowledge of concrete entities, actions, scenes and relations (such as those shown in Figure 3.8).

- The initial construction set consists of simple mappings linking word forms to embodied schemas. These include concrete referring expressions (for people, objects and places), action words and spatial relations (like the ones defined in Figure 4.3). No other function words (*e.g.*, determiners), relational constructions or other explicitly grammatical categories are present at the outset of learning.

- The initial construction set has a relatively flat organization. For example, no type distinction is initially made between common nouns and proper nouns, though this distinction could emerge through learning.

The inclusion of (some) simple lexical mappings in the initial grammar is not intended to suggest that all lexical learning precedes relational learning; in fact, the model assumes that both kinds of constructions can be learned based on the same general operations (see Section 6.2). But our current focus is on the particular challenges of learning early relational constructions; by this stage many lexical constructions will have been acquired by any of various learning strategies that have been proposed. This claim may be more reasonable for some conceptual domains (*e.g.*, object labels and names) than others (*e.g.*, verbs and spatial relations terms); but even inherently relational terms must be learned in part by association strategies; see Chapter 9 for further discussion of how and whether these assumptions can be relaxed.

Processes of language comprehension

The inclusion of language comprehension as part of the learner's prerequisite abilities follows naturally from the usage-based assumptions of our broader scientific framework. The simulation-based framework set forth in Chapter 3 proposes that language understanding is driven by active, context-sensitive mental simulations. A fully integrated model of (adult) language comprehension, including analysis, resolution and simulation, remains the subject of ongoing research. For the language learning stage under investigation, the simplified model of language understanding described in Chapter 4 is sufficient:

- The constructional analyzer takes input tokens (as described in Section 5.2.4) and an ECG grammar and produces constructional analyses and their corresponding semspecs. Analyses are ranked according to the scoring criteria that evaluate each of the constructional, form and meaning domains.

- The contextual resolver finds the best resolution map between a semspec and a given input context. No context history across utterances is assumed, and no explicit creation of referents (*e.g.*, based on indefinite referring expressions) is currently allowed. Resolution maps are ranked according to scoring criteria that rewards maps that have the greatest alignment between the semspec and context.

These language understanding processes satisfy the key requirements of the learning model: they must perform robustly with incomplete grammars; they must provide partial analyses and resolution maps that motivate the formation of new constructions; and they must provide a means of ranking and evaluating analyses.

5.2.4 Input data

The input to learning is characterized as a corpus of **input tokens**, each consisting of an utterance paired with its communicative context (as described in Section 4.1.2). The contents of both utterance and context are intended to reflect the sensorimotor, conceptual and inferential abilities of a human learner at the relevant stage, as reviewed in Chapter 2. The most relevant assumptions are as follows:

- The learner can pragmatically filter the glut of continuous percepts in the environment to extract those objects, relations and intentions most relevant to the attended scene and chunk them into discrete participants and events (*i.e.*, perform scene parsing).
- Context items instantiate schemas in the schema set S. The choice of S (as opposed to the ontology O) as the basis for contextual representation is consistent with the current focus on construction learning, as opposed to (linguistic) schema learning. The assumption is that linguistically relevant features may be more salient than others for the purposes of understanding and learning language.[5]
- The learner can reliably associate utterances with the appropriate scenes or referents, irrespective of their precise temporal alignment (*i.e.*, whether the utterance occurred before, during or after the event). In some cases, as in the imperative speech act of the example situation, the associated event may be inferred from context, whether or not the desired event ultimately takes place.
- The learner receives only positive examples, in the sense that tokens are not identified as non-occurring.

[5] This claim might be considered a comprehension-based analogue to Slobin's (1991) "thinking for speaking" hypothesis.

- No extended contextual history is represented in the model. A limited notion of shared situational context is available: each input token is associated with an *episode*, which has a persistent set of (shared) episode participants. In general, however, input tokens are treated by the analysis, resolution and learning processes as standalone units with direct access to any relevant contextual information.

Figure 5.1 shows an example input token similar to the earlier one in Figure 4.4. The same utterance "throw the ball" (with falling intonation) is paired with a context that explicitly represents the (desired) throwing schema requested of the child (Naomi) by the mother. The inclusion of a Throw context item reflects the assumptions above: namely, that the child can infer the mother's communicative intention from pragmatic cues (*e.g.*, gaze, gesture).

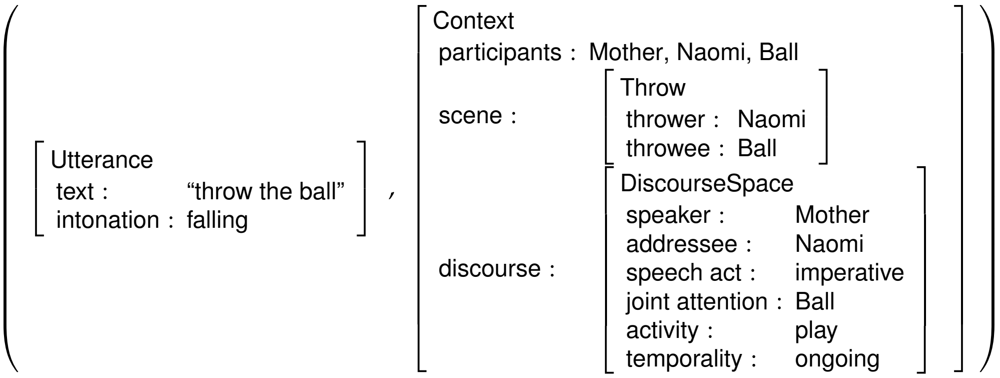

Figure 5.1. A typical input token for learning: the utterance "throw the ball" paired with a communicative context in which the Mother tells the child to throw the ball.

The specific training corpus used in learning experiments is a subset of the Sachs corpus of the CHILDES database of parent-child transcripts (Sachs 1983; MacWhinney 1991), annotated by developmental psychologists as part of a study of motion utterances (May *et al.* 1996). These annotations indicate semantic and pragmatic features available in the scene; they are described in more detail in Section 8.1.1.

5.2.5 Performance criteria

Under the assumption that communicative competence is a goal (if not necessarily the only goal) of language learning, the evaluation criteria used for measuring comprehension can also be exploited to gauge the overall progress of the learner. By testing the learning model at regular intervals during training, we can assess how new constructions incrementally improve its ability to comprehend new input tokens.

In general, any measure associated with a more broadly encompassing theory of language use could be incorporated into the grammar performance score, rewarding communicative success in both comprehension and production, or perhaps integrating an explicit notion of agent utility or goal achievement. For example, given a model of comprehension with access to a simulation engine, the learner's performance score could include a measure of the (contextually appropriate) inferences resulting from simulation. Likewise, a model capable of language production could include a measure of how well a produced utterance expresses a given communicative intention, or whether it succeeds in achieving a particular agent goal.

The performance criteria used for the current model are based on those defined in Section 4.3 for the analysis and resolution processes. Recall that those measures allow the analyzer and resolver to rank candidate analyses and resolution maps. When taken in aggregate (or averaged) across the input tokens of a test corpus, however, they also provide a measure that can be compared across grammars. The most relevant criteria are the following:

- *Token score:* The *token score* (*i.e.*, the resolved analysis score) serves as an overall indication of comprehension, since it aggregates constructional, form and meaning components resulting from both analysis and resolution.

- *Form-meaning coverage:* The *form-meaning score*, along with the domain-specific precision and recall scores on which it is based, provides a measure of how completely a grammar accounts for input utterances and contexts.

- *Analysis size:* The absolute size of analyses (*i.e.*, the number of form units and relations $|a_f|$ and the number of meaning schemas and bindings $|a_m| = |ss|$ accounted for) may increase as the grammars improve. Better grammars should also produce more complete analyses (*i.e.*, with a single spanning root construct) and a lower average number of roots.

The measures above provide a quantitative basis for evaluating grammar performance. As discussed further in Chapter 8, it is also useful to make qualitative assessments of the kinds of constructions learned and errors produced by the model, relative to observed patterns of child language acquisition.

5.2.6 Summary: relational construction learning

The specific problem faced by the learner in acquiring relational constructions can be summarized as follows:

Comprehension-based learning of relational constructions

hypothesis space The target of learning is a grammar $G = (S, C)$ defined using the ECG formalism, where S is a fixed schema set.

prior knowledge The learner has an initial grammar $G_0 = (S_0, C_0)$, with C_0 consisting of lexical constructions (for concrete objects, actions and relations) and a language understanding model that performs constructional analysis and contextual resolution.

input data The learner encounters a sequence of input tokens from a training corpus D, each represented as a pair of discrete feature structures describing the utterance form and its communicative context.

performance criteria The grammar performance of successive grammars G_i on a test corpus is measured by improvement in comprehension, using the criteria from the language understanding model including form-meaning coverage, analysis size, completeness and ambiguity.

This view of the core computational problem departs significantly from traditional framings of grammatical induction. Each aspect of the problem fundamentally incorporates meaning and use, and the choice of a construction-based grammar as the target of learning introduces structural complexity in each of the constructional, form and meaning domains. But the challenges of adopting such rich representations are more than offset by the advantages of including similarly rich structures corresponding to the learner's prior conceptual, linguistic and pragmatic knowledge. Moreover, the learner has access to comprehension processes with which to make sense of input utterances in context.

The remainder of this chapter explores solutions that exploit both of these factors — the richer representations available to the learner, and the language comprehension process itself — to overcome the greater inherent representational complexity of learning relational constructions.

5.3 Approaches to learning

The problem as set forth above presents a combination of challenges not directly addressed by previous approaches:

- The target of learning is an inventory of constructions on a continuum of abstraction and size, including both lexically specific and more general patterns, and both simple and structured mappings across the form and meaning domains.

- The space of possibilities is not pre-specified in terms of a fixed set of parameters.

- The goal of learning is to move toward grammars that allow progressively better comprehension of new input.

- The learning strategy must be incremental and sensitive to statistical regularities in the data.

Together these constraints preclude an exhaustive search through all possible grammars for a uniquely specified "correct" grammar. Rather, the topology of the space of grammars must be discovered (or constructed) on the basis of experience. But a purely bottom-up, instance-based approach will also not suffice, since some ability to generalize beyond seen data is required. We thus seek a learning algorithm that strategically converges on simpler (and more general) grammars without unduly sacrificing closeness of fit to observed data.

As suggested in Section 2.3.4, machine learning techniques in the closely related Bayesian and information-theoretic traditions provide the most natural candidates for capturing this tradeoff. The approach taken here can be seen as adapting previous work along these lines to accommodate the structural complexity of the target ECG representation and exploit the tight integration between language learning and understanding. After reviewing the basic Bayesian approach to model selection (Section 5.3.1), this section describes key features of the most relevant previous model of language learning, Bayesian model merging (Section 5.3.2) and then identifies the challenges involved in adapting these to the current task (Section 5.3.3).

5.3.1 Bayesian inference

Bayesian methods are by this point well-established both within the AI community and in applications that range across a swath of scientific realms. They are amply documented in both the literature (Russell & Norvig 1994; Jurafsky & Martin 2000); here I briefly review the probabilistic basis for Bayesian inference. In general, probability theory can be applied to assess the probability of a given event, and by extension the relative probabilities of various events; the event in question may also be a candidate hypothesis about some phenomenon.

In particular, the probabilistic event may be a hypothesis about model structure, where *model* refers in a general sense to any account for some observed data. That is, the goal is to select the most likely model M given the data D, *i.e.*, the one with the highest *conditional* probability $P(M|D)$. Though it is in principle possible to estimate $P(M|D)$ directly, it is sometimes more convenient to exploit the relation between the joint and conditional probability, expressed in (5.3.1), to produce the alternate expression for $P(M|D)$ in (5.3.2):

$$P(M \cap D) = P(M) \cdot P(D|M) = P(D) \cdot P(M|D) \tag{5.3.1}$$

$$P(M|D) = \frac{P(M) \cdot P(D|M)}{P(D)} \tag{5.3.2}$$

The Bayesian insight is that this rearrangement of the terms allows one to estimate $P(M|D)$ (which in this formulation is called the *posterior* probability) in terms of other quantities that are often more convenient to obtain, namely: the *prior* expectation $P(M)$ over what the model should be; the *likelihood* $P(D|M)$ that a given model M would give rise to the actual data D encountered; and a prior expectation $P(D)$ over the data itself.

The prior $P(M)$ is typically defined to encode biases dependent on the model space, often favoring simpler models over more complex models, and $P(D|M)$ is typically directly provided by the model. In choosing the best model $\hat{M}$ among all models M, one need not compute the denominator, $P(D)$, since it is the same for all candidate models:

$$\begin{aligned} \hat{M} &= \underset{M}{\text{argmax}} \; P(M|D) && (5.3.3) \\ &= \underset{M}{\text{argmax}} \; \frac{P(M) \cdot P(D|M)}{P(D)} && (5.3.4) \\ &= \underset{M}{\text{argmax}} \; P(M) \cdot P(D|M) \quad , && (5.3.5) \end{aligned}$$

where the $\underset{M}{\text{argmax}}$ operator selects the model M that maximizes the term that follows.

This *maximum a posteriori* (or *MAP*) estimate has found wide applicability in all domains of probabilistic reasoning; the model space M might range over a class of mathematical functions, a set of variables describing a world state, the hidden cause of a medical condition — indeed, any set of competing accounts of some observed data. The current discussion is motivated by the search for linguistic knowledge, where the model space is the space of grammars that can explain a body of linguistic experience.

5.3.2 Bayesian model merging

The closest antecedent to the current enterprise is Bayesian model merging. The basic idea behind model merging (Omohundro 1992) is to treat every encountered instance as an exemplar to be incorporated in its entirety into an (initially) unstructured overall model (in the general sense of *model* noted above). These initial submodels are thus maximally specific, so the model will perform very well on data previously observed (or similar to that previously observed), but poorly on data ranging further afield from its experience. But as more instances are incorporated, the algorithm incrementally modifies the model — in particular, by *merging* submodels — to reflect regularities in the data, as captured by similarities in the submodels. Every merge operation yields a more general model that accounts for a wider range of data.

As with other specific-to-general learning algorithms, the challenge is knowing when to stop:

stopping too early runs the risk of essentially memorizing the data, with the submodels still hewing to specific observed instances; stopping too late may lead to a vacuous model whose powers of discrimination have been compromised. Moreover, at any point, many candidate merge operations may be possible; some of these may strand the model in an inhospitable region of the search space, perhaps even a local minimum.

The Bayesian variant of model merging (Stolcke 1994; Stolcke & Omohundro 1994) addresses these problems by applying the Bayesian MAP criterion to control the merging process, that is, by selecting at every step the merge that leads to the largest increase in the overall probability of the model given the data encountered so far. The MAP estimate provides the means of guiding (and stopping) the search over possible merges. Each merge results in a model that is simpler (*i.e.*, has fewer models, and thus has a higher prior probability) but less specific to the data (*i.e.*, has lower likelihood). The algorithm chooses merges that increase the model's posterior probability, and stops when such merges are no longer available. Stated in its most general form:

Bayesian model merging algorithm (Stolcke 1994; Stolcke & Omohundro 1994)

1. **Data incorporation.** Given a set of examples D, build an initial model M_0 that explicitly includes each example.
2. **Structure merging.** Until the posterior probability decreases:
    (a) Find the candidate merge of substructures that results in the greatest increase in posterior probability.
    (b) Perform the candidate merge and remove the original structures.

The algorithm has a number of properties that make it a particularly appealing and cognitively plausible candidate for the current task. Models at all levels of specificity can happily coexist; aspects of both imitative learning (especially early in learning) and generalization (as more data is encountered) are reflected in model performance, just as in patterns of child language acquisition; and posterior probability provides a principled evaluation metric for guiding the search toward incrementally better models. The algorithm also lends itself to an online version, in which the two steps can be repeated for batches of data (or, in the limit, single examples).

Examples

The following two applications of model merging are especially relevant to the current problem:

Probabilistic attribute grammars. Stolcke (1994) applies model merging to learn several classes of probabilistic language models, including hidden Markov models, probabilistic context-free grammars and probabilistic attribute grammars. The last of these is the most relevant here: a *probabilistic attribute grammar* (PAG) extends a stochastic context-free grammar backbone with probabilistic feature constraints, which can express simple semantic relations. Input is drawn from the L_0 task mentioned in Section 1.2.2 (Feldman *et al.* 1996), based on simple scenes of shapes in a trajector-landmark configuration. Thus sentences like “a circle is above a square” are paired with attribute-value pairs describing the scene (*e.g.*, “tr=circle lm=square rel=above”). The model successfully learns PAGs that parse the input sentence and produce accurate corresponding scene descriptions (analogous to the task of comprehension).

Lexical semantic representations. Bailey (1997) applies the model merging algorithm to a semantically richer though syntactically simpler domain than Stolcke’s PAG domain to model the crosslinguistic acquisition of hand action verbs. The semantic domain draws on an active motor representation based on the *x-schema* formalism described in Section 2.3.3. The input consists of single words (*e.g.*, *push* and *shove*) paired with features structures parameterizing x-schemas of the associated actions (*e.g.*, “schema=push force=high posture=palm direction=away ...”). Merging these feature structures yields fewer structures with broader multinomial probability distributions over their features, resulting in a final lexicon containing one or more submodels for each verb. It thus exhibits both polysemy (*e.g.*, with different models for *push*ing a block and *pushing* a button) and near-synonymy (*e.g.*, similar models for *push* and *shove*, where the latter has a higher force component). The model also demonstrates how the same underlying action description, motivated by presumed universals of motor control, can give rise to crosslinguistic diversity in systems for naming actions, as shown for languages including Tamil, Farsi, Spanish and Russian. Learned lexicons allow the model to perform successfully on single-word versions of comprehension” (generation of a feature structure based on a verb) and production (selection of a verb given a feature structure).

These examples demonstrate how the general model merging algorithm can be adapted for a given domain, as summarized below in terms of four main components:

Data incorporation strategy: how to incorporate new data, that is, how to construct the initial model. Typically, a (sub)model is created for each input example, effectively memorizing the data encountered. The PAG case, for example, simply adds a new grammar rule expanding the start symbol into the input sentence and features; while the verb learning model adds a specific verb submodel matching the input verb and feature structure.

Structure merging operations: how to merge substructures. The merging operations typically perform generalization over submodels appropriate to each domain. The verb sense model combines probability distributions over each feature of two merged submodels for a given verb. The PAG model offers more structure (both CFG rules and feature equations) and correspondingly more merging operations. In addition to a generalization operator that merges rules with similar expansions to create more general rules, the model allows *chunking* (or composition) of two nonterminals into a separate new nonterminal, as well as additional operations on the feature equations.

Search strategy: how to search the model space for candidate merges. An exhaustive search through all possible pairs of structures may be possible for small domains, but in practice, heuristics for guiding the search are useful and necessary to reduce computation time. The PAG model and others described in Stolcke (1994) apply best-first search with some lookahead, while the verb sense model uses a *similarity* metric to keep track of the most similar feature structures as the best candidates for merging.

Evaluation criteria: how to compare competing candidates for merging, that is, how to measure the prior and likelihood used to compute posterior probability. In both cases above, the priors bias models toward simplicity (fewer and shorter rules; fewer total verb senses), while the likelihoods are calculated according to the relevant probabilistic model. The PAG model also keeps counts of rule use to avoid costly reanalysis of the entire input data.

Both of these examples are direct predecessors of the current model. The PAG formalism exhibits structural complexity: rewrite rules expand nonterminals to sequences that may include nonterminals that are themselves expanded. This complexity corresponds directly to ECG constituent structure. The verb learning model shows how semantically richer representations can be learned, subject to many of the same cognitive and developmental constraints as the present model.

5.3.3 Model merging for embodied constructions

How can the model merging paradigm be applied to the construction learning scenario at hand? As noted earlier, the domain of embodied constructions differs in important respects from previous language formalisms. A naive translation of the components above will therefore not suffice. Extensions to the model are prompted by three main considerations: the nature of the ECG representation itself; the tight relationship between language learning and understanding; and cognitive constraints on learning.

Representational expressivity. The ECG formalism has explicit notions of category structure, constituency and constraints. Additional structure comes from its basic function of linking the two domains of form and meaning. The model merging algorithm must be modified to handle these sources of structure, both within and across constructions. Operations for modifying the grammar must be extended to accommodate the internal structure of relational constructions, in particular to refine the notion of similarity to recognize the potential for shared, unshared and overlapping structure across constructions. In addition, appropriate evaluation criteria must be defined for scoring ECG grammars, corresponding to the Bayesian prior probability but suitable for the discrete nature of (this version of) the ECG formalism.

Usage-based learning. Processing considerations necessitate fundamental modifications to the model. Unlike previous model merging applications, in which training examples are initially incorporated directly (*i.e.*, memorized) as exemplars, the path from input token to construction is mediated by language comprehension. In fact, the goal of learning is to discover how existing structures can be exploited and combined into larger relational constructions. Thus, it is most economical to take advantage of the partial analyses provided by language comprehension, adding new constructions only as needed to supplement and extend existing constructions. Usage processes should also directly guide the search for grammar modifications: the choice of best operation may be triggered by specific instances of usage as they are analyzed and resolved. Finally, appropriate evaluation criteria must be defined for evaluating how well a grammar performs on a corpus, corresponding to the likelihood. The constructional score defined in Section 4.3 partially fulfills this function, since it measures the constructional probability $P(d|G)$ of the input token d given the grammar G. But the other factors included in Section 4.3 for measuring how well the grammar facilitates language comprehension should also be incorporated, as in Bryant (2008).

Cognitive constraints. Some additional changes to standard model merging are motivated by cognitive considerations, in particular the constraint that the learner may have limited storage memory capacity for both input tokens and constructions, as well as limited working memory to support the learning process. Incorporating each new example as a separate model, for example, may be reasonable for (some) lexical items and short multi-word sequences, but automatically codifying every new input token in its full glory as a novel construction seems less cognitively plausible. It may also be infeasible to search through the entire space of possible structure merging operations at every step, or to assess how a candidate construction affects a grammar's posterior probability by reevaluating the entire corpus of previously seen data. Appropriate heuristics for limiting the search space or reducing computation are needed; these will be discussed in Chapter 7..

5.4 Usage-based construction learning: overview

This section synthesizes the considerations above to present a usage-based model of construction learning that addresses the general class of language learning problems defined in Section 5.2. Figure 5.2 revisits the integrated understanding-learning cycle introduced in Chapter 1 (Figure 1.1). The structures and processes involved in the language understanding pathway — including embodied schemas and constructions, the input token of an utterance in its situational context, the analysis and resolution processes and the resulting semspec — have now been explicitly defined.

The language understanding process provides a direct usage-based impetus for the language learning part of the cycle, depicted here as two processes: the *hypothesis* of new constructions and the *reorganization* of existing ones. These construction learning operations, corresponding to the structure merging operations in model merging, are motivated and constrained by the nature of the target space, the processes of language comprehension and cognitive considerations. They are also are mediated by quantitative evaluation metrics (not shown in the figure), approximating Bayesian scoring criteria. As the learner encounters more data in the usage-learning loop, the constructional mappings learned and refined should facilitate increasingly more complete and accurate comprehension of new data.

Figure 5.3 gives a high-level incremental learning algorithm corresponding to Figure 5.2, along with a version taking a corpus of input tokens that simply loops over the above procedure. The algorithm provides a class of solutions to the language learning problems defined in Section 5.2, adapting Bayesian model merging for the construction-learning domain to address the concerns outlined in Section 5.3.3 as follows:

Data incorporation: Input tokens are not incorporated directly as new constructions, but are instead first analyzed and resolved using the current grammar. They are indirectly incorporated via learning operations that propose new constructions based on the results of analysis.

Grammar update operations: Structure merging operations are extended to allow both the hypothesis of new constructions and the reorganization of existing ones. These must satisfy both the structural demands of the ECG construction formalism (to exploit the presence of shared internal structure) and the process-based demands of language analysis (to improve comprehension of new input).

Search strategy: The search for candidate constructions is guided by incoming data. In some cases, the results of analyzing the current input token directly triggers specific learning operations.

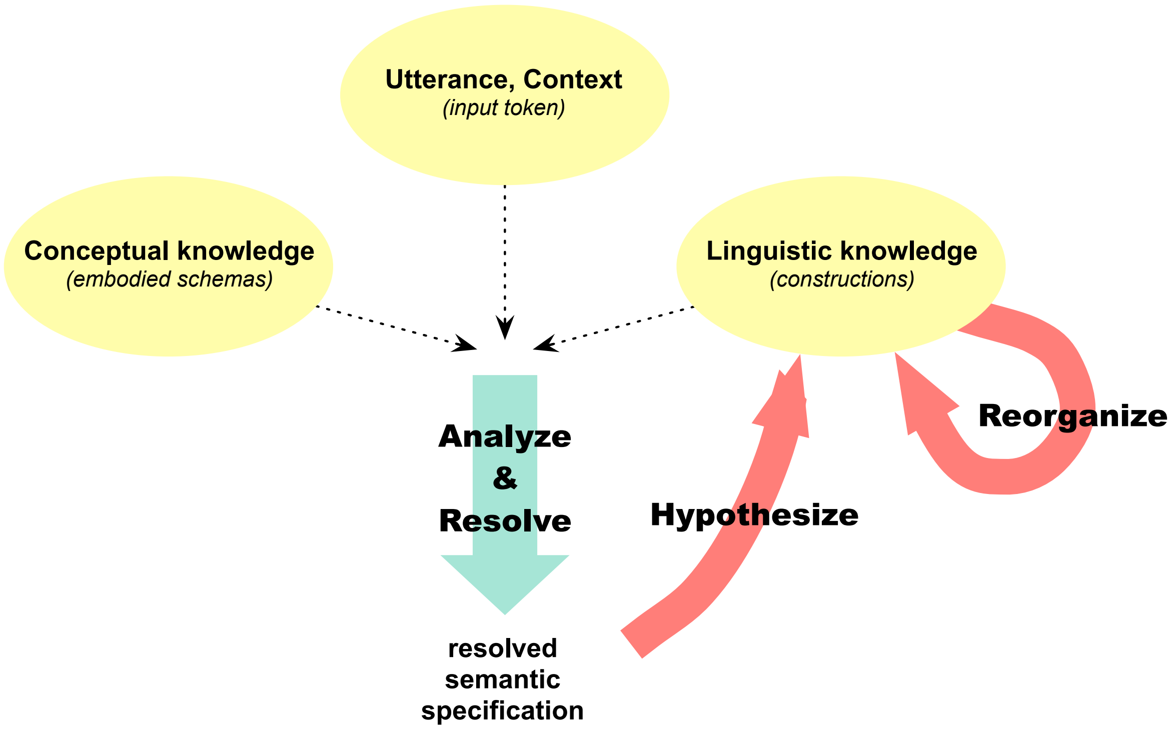

Figure 5.2. Usage-based construction learning: Input tokens (utterance-context pairs) are analyzed and resolved, producing a resolved semantic specification and prompting the hypothesis of new constructions and the reorganization of existing ones.

```
Learn from input token: Given input token $d$ and grammar $G$, return grammar $G'$.

1. Analyze $d$ using $G$, producing analysis $a_d$ and resolved semspec $rss_a$.
2. Find the set $L$ of candidate learning operations based on $a_d$, $rss_a$ and $G$, by hypothesizing
   new constructions and reorganizing existing constructions.
3. For each learning operation $l \in L$:
    (a) Create new grammar $G_l$ by performing operation $l$ on $G$.
    (b) Calculate the grammar improvement $\Delta(G_l, G)$.
4. Set $G'$ to the $G_l$ that maximizes $\Delta(G_l, G)$.
5. If $\Delta(G', G)$ exceeds threshold of improvement, return $G'$; else return $G$.
```

```
Learn from corpus: Given a corpus $D$ and a grammar $G_0$, return grammar $G_1$.

1. Initialize current grammar $G$ to $G_0$.
2. For each input token $d \in D$, learn from $d$ and current grammar $G$, producing $G'$.
3. Return current grammar as $G_1$.
```

Figure 5.3. Usage-based construction learning algorithm: algorithms for learning new grammars from a single token and from a corpus.

In others, the search for new constructions may be restricted to a subset of the entire repository (*e.g.*, recently used constructions and their nearest neighbors in the construction set).

Evaluation criteria: The grammar improvement $\Delta(G', G)$ corresponds to the posterior probability used in model merging, but it is adapted for the discrete structures of the ECG domain to use a simplicity-based criterion that balances a prior favoring simpler grammars against a likelihood favoring simpler analyses of the data using the grammar (*i.e.*, encoding a bias toward better prediction of the data). The calculation of this score may be restricted to a subset of the corpus (*e.g.*, recently encountered tokens, or especially relevant or problematic tokens).

Besides the representational modifications needed for the ECG domain, the most significant adaptations to each component above are motivated by aspects of language comprehension. In particular, the shift to a data-driven, analyzer-mediated basis for processing input removes the strict boundary between data incorporation and the search for new constructions: both depend on the ongoing flow of data as it is processed by the current grammar. The model is thus usage-based in that its path through the space of grammars depends directly on specific instances of language use. It is also usage-based in the related sense of exploiting the statistical characteristics of usage over time, as reflected in the model's scoring criterion.

Like the general class of problems it addresses, this class of solutions admits many instantiations; each of the algorithmic components above leaves significant room for interpretation. The next two chapters together instantiate a solution that satisfies the constraints of the construction learning problem defined in this chapter, where the ECG-based structures described in Chapter 3 and, especially, the language analysis processes described in Chapter 4 bear directly on every aspect of the model.

- Chapter 6 addresses issues related to the search for candidate constructions, motivated by both the nature of the search space and domain-specific strategies suggested by developmental evidence. I define several operations for the hypothesis and reorganization of constructions and describe the conditions under which they apply (corresponding to step 2 of the high-level algorithm).

- Chapter 7 focuses on the quantitative evaluation of candidate grammars, defining a heuristic for calculating the grammar improvement $\Delta(G', G)$ based on minimum description length. This heuristic is guided by an information-theoretic bias toward a minimal encoding of the grammar together with the data.

In fleshing out the requisite details, I will make a number of simplifying assumptions to ease both exposition and implementation. It is essential to bear in mind that the intent is not to define an exhaustive set of learning operations or the most statistically sensitive evaluation criteria possible. The nature of the problem space and relative paucity of data available for experimentation make it unlikely to reward much ingenuity in model design at this initial stage. Fortunately, the learning framework guarantees that with enough data, any reasonable optimization criteria should lead to improvement over time. The goal here is thus to define a basic toolkit of usage-based operations and criteria, sufficient to demonstrate how the underlying dynamics of the model push it toward increasingly better grammars over the course of experience.