2

A Variational Model of Language Acquisition

One hundred years without Darwin are enough.
H. J. Muller (1959), on the centennial of *On the Origin of Species*

It is a simple observation that young children's language is different from that of adults. However, this simple observation raises profound questions: What results in the differences between child language and adult language, and how does the child eventually resolve such differences through exposure to linguistic evidence?

These questions are fundamental to language acquisition research. (6) in Chapter 1, repeated below as (14), provides a useful framework within to characterize approaches to language acquisition:

(14) $\mathcal{L}$: $(S_0, E) \rightarrow S_T$

Language acquisition can be viewed as a function or algorithm, $\mathcal{L}$, which maps the initial and hence putatively innate state (S_0) of the learner to the terminal state (S_T), the adult-form language, on the basis of experience, E, in the environment.

Two leading approaches to $\mathcal{L}$ can be distinguished in this formulation according to the degree of focus on S_0 and $\mathcal{L}$. An empiricist approach minimizes the role of S_0, the learner's initial (innate) and domain-specific knowledge of natural language. Rather, emphasis is given to $\mathcal{L}$, which is claimed to be a generalized learning mechanism cross-cutting cognitive domains. Models in this approach can broadly be labeled *generalized statistical learning* (GSL): learning is the approximation of the terminal state (S_T) based on the statistical distribution of the input data. In contrast, a rationalist approach, often rooted in the tradition of generative grammar, attributes the success of language acquisition to a richly endowed S_0, while relegating $\mathcal{L}$ to a background role. Specifically, S_0 is assumed to be a delimited space, a Universal Grammar (UG), which consists of a finite number of hypotheses that a child can in principle entertain. Almost all theories of acquisition in the UG-based approach can called *transformational learning* models, borrowing a term from evolutionary biology (Lewontin 1983): the learner's linguistic hypothesis undergoes direct transformations (changes), by moving from one hypothesis to another, driven by linguistic evidence.

This study introduces a new approach to language acquisition in which both S_0 and $\mathcal{L}$ are given prominent roles in explaining child language. We will show that once the domain-specific and innate knowledge of language (S_0) is assumed, the mechanism language acquisition ($\mathcal{L}$) can be related harmoniously to the learning theories from traditional psychology, and possibly, the development of neural systems.

2.1 Against transformational learning

Recall from Chapter 1 the three conditions on an adequate acquisition model:

(15) a. formal sufficiency
b. developmental compatibility
c. explanatory continuity

If one accepts these as guidelines for acquisition research, we can put the empiricist GSL models and the UG-based transformational learning models to the test.

In recent years, the GSL approach to language acquisition has (re)gained popularity in cognitive sciences and computational linguistics (see e.g. Bates & Elman 1996, Seidenberg 1997). The GSL approach claims to assume little about the learner's initial knowledge of language. The child learner is viewed as a generalized data

processor, such as an artificial neural network, which approximates the adult language based on the statistical distribution of the input data. The GSL approach claims support (Bates & Elman 1996) from experiments showing that infants are capable of extracting statistical regularities in (quasi)linguistic information (e.g. Saffran et al. 1996).

Despite this renewed enthusiasm, it is regrettable that the GSL approach has not tackled the problem of language acquisition in a broad empirical context. For example, a main line of work (e.g. Elman 1990, 1991) is dedicated to showing that certain neural network models are able to capture some limited aspects of syntactic structures—a most rudimentary form of the formal sufficiency condition—although there is still debate on whether this project has been successful (e.g. Marcus 1998). Much more effort has gone into the learning of irregular verbs, starting with Rumelhart & McClelland (1986) and followed by numerous others,[1] which prompted a review of the connectionist manifesto, *Rethinking Innateness* (Elman et al. 1996), to remark that connectionist modeling makes one feel as if developmental psycholinguistics is only about 'development of the lexicon and past tense verb morphology'(Rispoli 1999: 220). But even for such a trivial problem, no connectionist network has passed the *Wug*-test (Prasada & Pinker 1993, Pinker 1999), and, as we shall see in Chapter 3, much of the complexity in past tense acquisition is not covered by these works.

As suggested in section 1.2.2, there is reason to believe that these challenges are formidable for generalized learning models such as an artificial neural network. Given the power of computational tools available today, it would not be remarkable to construct a (GSL) system that learns something. What *would* be remarkable is to discover whether the constructed system learns in much the same way that human children learn. (10) shows that child language and adult language display significant disparities in statistical distributions; what the GSL approach has to do, then, is

[1] Pinker (1999: 302) lists 25 major connectionist studies on irregular verbs.

to find an empiricist (learning-theoretic) alternative to the learning biases introduced by innate UG. This seems difficult, given the simultaneous constraints—from both child language acquisition and comparative studies of the world's languages—that such an alternative must satisfy. That is, an empiricist must account for, say, systematic utterances like *me riding horse* (meaning 'I am riding a horse') in child language and island constraints in adult language, at the same time. But again, nothing can be said unless the GSL approach faces the challenges from the quantitative and crosslinguistic study of child language; as pointed out by Lightfoot (1998), Fodor & Crowther (in press), and others, there is nothing on offer.

We thus focus our attention on the other leading approach to language acquisition, which is most closely associated with generative linguistics. We will not review the argument for innate linguistic knowledge; see section 1.1 for a simple yet convincing example. The restrictiveness in the child language learner's hypothesis space, coupled with the similarities revealed in comparative studies of the world's languages, have led linguists to conclude that human languages are delimited in a finite space of possibilities, the Universal Grammar. The Principles and Parameters (P&P) approach (Chomsky 1981) is an influential instantiation of this idea by attempting to constrain the space of linguistic variation to a set of parametric choices.

In generative linguistics, the dominant model of language acquisition (e.g. Chomsky 1965, Wexler & Culicover 1980, Berwick 1985, Hyams 1986, Dresher & Kaye 1990, Gibson & Wexler 1994) can be called the *transformational learning* (TL) approach. It assumes that the state of the learner undergoes direct changes, as the old hypothesis is replaced by a new hypothesis. In the *Aspects*-style framework (Chomsky 1965), it is assumed (Wexler & Culicover 1980, Berwick 1985) that when presented with a sentence that the learner is unable to analyze with the present set of rules, an appropriate rule is added to the current hypothesis. Hence, a new hypothesis is formed to replace the old. With the advent of the P&P framework, acquiring a language has been

viewed as setting the appropriate parameters. An influential way to implement parameter setting is the *triggering* model (Chomsky 1981, Gibson & Wexler 1994). In a typical triggering algorithm, the learner changes the value of a parameter in the present grammar if the present grammar cannot analyze an incoming sentence and the grammar with the changed parameter value can. Again, a new hypothesis replaces the old hypothesis. Note that in all TL models, the learner changes hypotheses in an all-or-nothing manner; specifically for the triggering model, the UG-defined parameters are literally 'triggered' (switched on and off) by the relevant evidence. For the rest of our discussion, we will focus on the triggering model (Gibson & Wexler 1994), representative of the TL models in the UG-based approach to language acquisition.

2.1.1 *Formal insufficiency of the triggering model*

It is by now well known that Gibson & Wexler's triggering model has a number of formal problems (see Berwick & Niyogi 1996, Frank & Kapur 1996, Dresher 1999). The first problem concerns the existence of local maxima in the learning space. Local maxima are non-target grammars from which the learner can never reach the target grammar.[2] By analyzing the triggering model as a Markovian process in a finite space of grammars, Berwick & Niyogi (1996) have demonstrated the pervasiveness of local maxima in Gibson and Wexler's (very small) three-parameter space. Gibson & Wexler (1994) suggest that the local maxima problem might be circumvented if the learner starts from a default parameter setting, a 'safe' state, such that no local maximum can ever be encountered. However, Kohl (1999), using an exhaustive search in a computer implementation of the triggering model, shows that in a linguistically realistic twelve-parameter space, 2,336 of the 4,096 grammars are still not learnable even

[2] The present discussion concerns acquisition in a *homogeneous* environment in which all input data can be identified with a single, idealized 'grammar'. For historical reasons we continue to refer to it by the traditional term 'target grammar'.

with the best default starting state. With the worst starting state, 3,892 grammars are unlearnable. Overall, there are on average 3,348 unlearnable grammars for the triggering model.[3]

A second and related problem has to do with the ambiguity of input evidence. In a broad sense, ambiguous evidence refers to sentences that are compatible with more than one grammar. For example, a sentence with an overt thematic subject is ambiguous between an English-type grammar, which obligatorily uses subjects, and a Chinese-type grammar, which optionally uses subjects. When ambiguous evidence is presented, it may select any of the grammars compatible with the evidence and may subsequently be led to local maxima and unlearnability. To resolve the ambiguity problem, Fodor's (1998) Structural Trigger Learner (STL) model assumes that the learner can determine whether an input sentence is unambiguous by attempting to analyze it with multiple grammars. Only evidence that unambiguously determines the target grammar triggers the learner to change parameter values. Although Fodor shows that there is unambiguous evidence for each of the eight grammars in Gibson & Wexler's three-parameter space, such optimistic expectations may not hold for a large parametric space in general (Clark 1992, Clark & Roberts 1993; we return to this with a concrete example in section 2.3.3). Without unambiguous evidence, Fodor's revised triggering model will not work.

Lastly, the robustness of the triggering model has been called into question. As pointed out by Osherson et al. (1984), Randall (1990), and Valian (1990), even a small amount of noise can lead the triggering-like transformational models to converge on a wrong grammar. In a most extreme form, if the *last* sentence the

[3] Niyogi & Berwick (1995) argue that 'mis-convergence', i.e. the learner attaining a grammar that is different from target grammar, is what makes language change possible: hence formal insufficiency of the triggering model may be a virtue instead of a defect. However, empirical facts from diachronic studies suggest a different picture of how language changes; see Ch. 5. In addition, whatever positive implications of misconvergence are surely negated by the overwhelming failure to converge, as Kohl's results show.

learner hears just before language acquisition stops happens to be noise, the learning experience during the entire period of language acquisition is wasted. This scenario is by no means an exaggeration when a realistic learning environment is taken into account. Actual linguistic environments are hardly uniform with respect to a single idealized grammar. For example, Weinreich et al. (1968: 101) observe that it is unrealistic to study language as a 'homogeneous object', and that the 'nativelike command of heterogeneous structures is not a matter of multidialectalism or "mere" performance, but is part of unilingual linguistic competence'. To take a concrete example, consider again the acquisition of subject use. English speakers, who in general use overt subjects, do occasionally omit them in informal speech, e.g. *Seems good to me.* This pattern, of course, is compatible with an optional subject grammar. Now recall that a triggering learner can alter its hypothesis on the basic of a *single* sentence. Consequently, variability in linguistic evidence, however sparse, may still lead a triggering learner to swing back and forth between grammars like a pendulum.

2.1.2 *Developmental incompatibility of the triggering model*

While it might be possible to salvage the triggering model to meet the formal sufficiency condition (e.g. via a random-walk algorithm of Niyogi & Berwick 1996; but cf. Sakas & Fodor 2001), the difficulty posed by the developmental compatibility condition is far more serious. In the triggering model, and in fact in *all* TL models, the learner at any one time is identified with a single grammar. If such models are at all relevant to the explanation of child language, the following predictions are inevitable:

(16) a. The learner's linguistic production ought to be consistent with respect to the grammar that is currently assumed.
b. As the learner moves from grammar to grammar, abrupt changes in linguistic expressions should be observed.

To the best of my knowledge, there is in general no developmental evidence in support of either (16a) or (16b).

A good test case is again children's null subjects (NS), where we have a large body of quantitative and crosslinguistic data. First, consider the prediction in (16a), the consistency of child language with respect to a single grammar defined in the UG space. Working in the P&P framework, Hyams (1986), in her groundbreaking work, suggests that English child NS results from missetting their language to an optional-subject grammar such as Italian, in which subject drop is grammatical. However, Valian (1991) shows that while Italian children drop subjects in 70% of all sentences, the NS ratio is only 31% for American children in the same age group. This statistical difference renders it unlikely that English children initially use an Italian-type grammar. Alternatively, Hyams (1991) suggests that during the NS stage, English children use a discourse-based, optional-subject grammar like Chinese. However, Wang et al. (1992) show that while subject drop rate is only 26% for American children during the NS stage (2;0–3;0),[4] Chinese children in the same age group drop subjects in 55% of all sentences. Furthermore, if English children did indeed use a Chinese-type grammar, one predicts that object drop, grammatical in Chinese, should also be robustly attested (see section 4.3.2 for additional discussion). This is again incorrect: Wang et al. (1992) find that for 2-year-olds, Chinese children drop objects in 20% of sentences containing objects and American children only 8%. These comparative studies conclusively demonstrate that subject drop in child English cannot be identified with any single adult grammar.

Turning now to the triggering models' second prediction for language development (16b), we expect to observe abrupt changes

[4] This figure, as well as Valian's (1991), is lower than those reported elsewhere in the literature, e.g. Bloom (1993), Hyams & Wexler (1993). However, there is good reason to believe that around 30% is a more accurate estimate of children's NS rate. In particular, Wang et al. (1992) excluded children's NS sentences such as infinitives and gerunds that would be acceptable in adult English; see Phillips (1995) for an extended discussion on the counting procedure.

in child language as the learner switches from one grammar to another. However, Bloom (1993) found no sharp changes in the frequency of subject use throughout the NS stage of Adam and Eve, two American children studied by Brown (1973). Behrens (1993) reports similar findings in a large longitudinal study of German children's NS stage. Hence, there is no evidence for a radical reorganization—parameter resetting (Hyams & Wexler 1993)—of the learner's grammar. In section 4.1 we will show that for Dutch acquisition, the percentage of V2 use in matrix sentences also rises gradually, from about 50% at 2;4 to 85% at 3;0. Again, there is no indication of a radical change in the child's grammar, contrary to what the triggering model entails. Overall, the gradualness of language development is unexpected in the view of all-or-none parameter setting, and has been a major argument against the parameter-setting model of language acquisition (Valian 1990, 1991, Bloom 1990, 1993), forcing many researchers to the conclusion that child and adult language differ not in competence but in performance.

2.1.3 *Imperfection in child language?*

So the challenge remains: what explains the differences between child and adult languages? As summarized in Chapter 1 and repeated below, two approaches have been advanced to account for the differences between child and adult languages:

(17) a. Children and adults differ in linguistic performance.
b. Children and adults differ in grammatical competence.

The performance deficit approach (17a) is often stated under the Continuity Hypothesis (Macnamara 1982, Pinker 1984). It assumes an identity relation between child and adult competence, while attributing differences between child and adult linguistic forms to performance factors inherent in production, and (nonlinguistic) perceptual and cognitive capacities that are still underdeveloped at a young age (e.g. Pinker 1984, Bloom 1990, 1993, Gerken 1991, Valian 1991).

The competence deficit approach (17b) is more often found in works in the parameter-setting framework. In recent years it has been claimed (Hyams 1996, Wexler 1998), in contrast to earlier ideas of parameter mis-setting, that the parameter values are set correctly by children very early on.[5] The differences between child language and adult language have been attributed to other deficits in children's grammatical competence. For example, one influential approach to the OI phenomenon reviewed in section 1.2.2 assumes a deficit in the Tense/Agreement node in children's syntactic representation (Wexler 1994): the Tense/Agreement features are missing in young children during the ROI stage. Another influential proposal in Rizzi's (1994) *Truncation Hypothesis* holds that certain projections in the syntactic representation, specifically CP, are missing in young children's knowledge of language. The reader is referred to Phillips (1995) for a review and critique of some recent proposals along these lines.

Despite the differences between the two approaches, a common theme can be identified: child language is assumed to be an *imperfect* form of adult language, perturbed by either competence or performance factors. In section 1.2.3, we have already noted some methodological pitfalls associated with such explanatorily discontinuous accounts. More empirically, as we shall see in Chapters 3 and 4, the imperfection perspective on child language leaves many developmental patterns unexplained. To give a quick preview, we will see that children's overregularization errors (*hold-holded*) reveal important clues on how phonology is structured and learned, and should not be regarded as simple memory retrieval failures as in Pinker (1999). We will see that when English children drop subjects in *Wh* questions, they do so almost always in adjunct (*where, how*) questions, but almost never in argument (*who, what*) questions: a categorical asymmetry not predicted by any imperfection explanation proposed so far. We will document the robust use

[5] Although it is not clear how parameters are set (correctly), given the formal insufficiency of the triggering model reviewed earlier.

(approximately 50%) of V1 patterns in children acquiring V2: hence, 50% of 'imperfection' to be explained away.

This concludes our very brief review of the leading approaches to language acquisition. While there is no doubt that innate UG knowledge must play a crucial role in constraining the child's hypothesis space and the learning process, there is *one* component in the GSL approach that is too sensible to dismiss. That is, statistical learning seems most naturally suited to modeling the gradualness of language development. In the rest of this chapter we propose a new approach that incorporates this useful aspect of the GSL model into a generative framework: an innate UG provides the *hypothesis space* and statistical learning provides the *mechanism*. To do this, we draw inspiration from Darwinian evolutionary biology.

2.2 The variational approach to language acquisition

2.2.1 *The dynamics of Darwinian evolution*

We started the discussion of child language by noting the variation between child and adult languages. It is a fundamental question how such variation is interpreted in a theory of language acquisition. Here, the conceptual foundation of Darwinian evolutionary thinking provides an informative lesson.

Variation, as an intrinsic fact of life, can be observed at many levels of biological organizations, often manifested in physiological, developmental, and ecological characteristics. However, variation among individuals in a population was not fully recognized until Darwin's day. As pointed out by Ernst Mayr on many occasions (in particular, 1963, 1982, 1993), it was Darwin who first realized that the variations among individuals are 'real': individuals in a population are inherently different, and are not mere 'imperfect' deviations from some idealized archetype.

Once the reality of variation and the uniqueness of individuals

were recognized, the correct conception of evolution became possible: variations at the individual level result in fitness variations at the population level, thus allowing evolutionary forces such as natural selection to operate. As R. C. Lewontin remarks, evolutionary changes are hence changes in the *distribution* of different individuals in the population:

> Before Darwin, theories of historical change were all *transformational*. That is, systems were seen as undergoing change in time because each element in the system underwent an individual transformation during its history. Lamarck's theory of evolution was transformational in regarding species as changing because each individual organism within the species underwent the same change. Through inner will and striving, an organism would change its nature, and that change in nature would be transmitted to its offspring.
>
> In contrast, Darwin proposed a *variational* principle, that individual members of the ensemble differ from each other in some properties and that the system evolves by changes in the proportions of the different types. There is a sorting-out process in which some variant types persist while others disappear, so the nature of the ensemble as a whole changes without any successive changes in the individual members. (Lewontin 1983: 65–6; italics original.)

For scientific observations, the message embedded in Darwinian variational thinking is profound. Non-uniformity in a sample of data often should, as in evolution, be interpreted as a collection of *distinct* individuals: variations are therefore real and expected, and should not be viewed as 'imperfect' forms of a single archetype. In the case of language acquisition, the differences between child and adult languages may not be the child's imperfect grasp of adult language; rather, they may actually reflect a principled grammatical system in development and transition, before the terminal state is established. Similarly, the distinction between transformational and variational thinking in evolutionary biology is also instructive for constructing a formal model of language acquisition. Transformational learning models identify the learner with a single hypothesis, which directly changes as input is processed. In contrast, we may consider a variational theory in which language acquisition is the change in the *distribution* of I-language grammars, the principled variations in human language.

In what follows, we present a learning model that instantiates the variational approach to language acquisition. The computational properties of the model will then be discussed in the context of the formal sufficiency condition on acquisition theories.

2.2.2 *Language acquisition as grammar competition*

To explain the non-uniformity and the gradualness in child language, we explicitly introduce statistical notions into our learning model. We adopt the P&P framework, i.e. assuming that there is only a finite number of possible human grammars, varying along some parametric dimensions. We also adopt the strongest version of continuity hypothesis, which says, without evidence to the contrary, that UG-defined grammars are accessible to the learner from the start.

Each grammar G_i is paired with a weight p_i, which can be viewed as the measure of prominence of G_i in the learner's language faculty. In a linguistic environment E, the weight $p_i(E, t)$ is determined by the learning function $\mathcal{L}$, the linguistic evidence in E, and the time variable t, the time since the outset of language acquisition. Learning stops when the weights of all grammars are stabilized and do not change any further,[6] possibly corresponding to some kind of critical period of development. In particular, in an idealized environment where *all* linguistic expressions are generated by a 'target' grammar T—again, keeping to the traditional terminology—we say that learning *converges to target* if $p_T = 1$ when learning stops. That is, the target grammar has eliminated all other grammars in the population as a result of learning.

The learning model is schematically shown below:

(18) Upon the presentation of an input datum s, the child
  a. selects a grammar G_i with the probability p_i
  b. analyzes s with G_i
  c. • if successful, reward G_i by increasing p_i
     • otherwise, punish G_i by decreasing p_i

[6] This does not mean that learning necessarily converges to a single grammar; see (24) below.

Metaphorically speaking, the learning hypotheses—the grammars defined by UG—*compete*: grammars that succeed in analyzing a sentence are rewarded and those that fail are punished. As learning proceeds, grammars that have overall more success with the data will be more prominently represented in the learner's hypothesis space.

An example illustrates how the model works. Imagine the learner has two grammars, G_1, the target grammar used in the environment, and G_2, the competitor, with associated weights of p_1 and p_2 respectively. Initially, the two grammars are undifferentiated, i.e. with comparable weights. The learner will then have comparable probabilities of selecting the grammars for both input analysis and sentence production, following the null hypothesis that there is a single grammatical system responsible for both comprehension/learning and production. At this time, sentence sequences produced by the learner will look like this:

(19) Early in acquisition:
$S_{G1}, S_{G1}, S_{G2}, S_{G1}, S_{G2}, S_{G2}, \ldots$

where S_G indicates a sentence produced by the grammar G.[7]

As learning proceeds, G_2, which by assumption is incompatible with at least *some* input data, will be punished and its weight will gradually decrease. At this stage of acquisition, sequences produced by the learner will look like this:

(20) Intermediate in acquisition:
$S_{G1}, S_{G1}, S_{G2}, S_{G1}, S_{G1}, S_{G1} \ldots$

where G_1 will be more and more dominantly represented.

When learning stops, G_2 will have been eliminated ($p_2 \approx 0$) and G_1 is the only grammar the learner has access to:

(21) Completion of acquisition:
$S_{G1}, S_{G1}, S_{G1}, S_{G1}, S_{G1}, S_{G1}, \ldots$

[7] It is possible that some sentences are ambiguous between G_1 and G_2, which may extensionally overlap.

Of course, grammars do not actually compete with each other: the competition metaphor only serves to illustrate (a) the grammars' coexistence and (b) their differential representation in the learner's language faculty. Neither does the learner play God by supervising the competition of the grammars and selecting the winners.[8] We will also stress the *passiveness* of the learner in the learning process, conforming to the research strategy of a 'dumb' learner in language acquisition. That is, one does not want to endow the learner with too much computational power or too much of an active role in learning. The justification for this minimum assumption is twofold. On the one hand, successful language acquisition is possible, barring pathological cases, irrespective of 'general intelligence'; on the other, we simply don't have a theory of children's cognitive/computational capacities to put into a rigorous model of acquisition—an argument from ignorance. Hence, we assume that the learner does not contemplate which grammar to use when an input datum is presented. He uses whichever happens to be selected with its associated weight/probability. He does not make active changes to the selected grammar (as in the triggering model), or reorganize his grammar space, but simply updates the weight of the grammar selected and moves on.

Some notations. Write $s \in E$ if a sentence s is an utterance in the linguistic environment E. We assume that during the time frame of language acquisition, E is a fixed environment, from which s is drawn independently. Write $G \rightarrow s$ if a grammar G can analyze s, which, as a special case, can be interpreted as parsability (Wexler & Culicover 1980, Berwick 1985), in the sense of *strong generative capacity*. Clearly, the weak generative notion of string-grammar acceptance does not affect formal properties of the model. However, as we shall see in Chapter 4, children use their morphological knowledge and domain-specific knowledge of UG—strong generative notions—to disambiguate grammars. It is worth noting that the formal properties of the model are independent of the definition of analyzability: any well-defined and empirically justified notion will suffice. Our choice of string-grammar compatibility obviously eases the evaluation of grammars using linguistic corpora.

[8] In this respect, the variational model differs from a similar model of acquisition (Clark 1992), in which the learner is viewed as a genetic algorithm that explicitly evaluates grammar fitness. We return to this in section 2.5.

Suppose that there are altogether N grammars in the population. For simplicity, write p_i for $p_i(E, t)$ at time t, and p_i' for $p_i(E, t+1)$ at time $t+1$. Each time instance denotes the presentation of an input sentence. In the present model, learning is the adaptive change in the weights of grammars in response to the sentences successively presented to the learner. There are many possible instantiations of competition-based learning.[9] Consider the one in (22):

(22) Given an input sentence s, the learner selects a grammar G_i with probability p_i:

a. if $G_i \rightarrow s$ then $\begin{cases} p_i' = p_i + \gamma(1 - p_i) \\ p_j' = (1-\gamma)p_j \quad \text{if } j \neq i \end{cases}$

b. if $G_i \not\rightarrow s$ then $\begin{cases} p_i' = (1-\gamma)p_i \\ p_j' = \dfrac{\gamma}{N-1} + (1-\gamma)p_j \quad \text{if } j \neq i \end{cases}$

(22) is the Linear reward-penalty (L_{R-P}) scheme (Bush & Mosteller 1951, 1958), one of the earliest, simplest, and most extensively studied learning models in mathematical psychology. Many similar competition-based models have been formally and experimentally studied, and receive considerable support from human and and animal learning and decision-making; see Atkinson et al. (1965) for a review.

Does the employment of a general-purpose learning model from the behaviorist tradition, the L_{R-P}, signal a return to the Dark Ages? Absolutely not. In competition learning models, what is crucial is the constitution of the hypothesis space. In the original L_{R-P} scheme, the hypothesis space consists of simple responses

[9] See Yang & Gutmann (1999) for a model that uses a Hebbian style of update rules.

conditioned on external stimulus; in the grammar competition model, the hypothesis space consists of Universal Grammar, a highly constrained and finite range of possibilities. In addition, as discussed in Chapter 1, it seems unlikely that language acquisition can be equated to data-driven learning without prior knowledge. And, as will be discussed in later chapters in addition to numerous other studies in language acquisition, in order adequately to account for child language development, one needs to make reference to specific characterization of UG supplied by linguistic theories.

There is yet another reason for having an explicit account of the learning process: because language *is* acquired, and thus the composition, distribution, and other properties of the input evidence, in principle, matter. The landmark study of Newport et al. (1977) is best remembered for debunking the necessity of the so-called 'Motherese' for language acquisition, but it also shows that the development of *some* aspects of language does correlate with the abundance of linguistic data. Specifically, children who are exposed to more yes/no questions tend to use auxiliary verbs faster and better. An explicit model of learning that incorporates the role of input evidence may tell us why such correlations exist in some cases, but not others (e.g. the null subject phenomenon). The reason, as we shall see, lies in the Universal Grammar.

Hence, our emphasis on $\mathcal{L}$ is simply a plea to pay attention to the actual mechanism of language development, and a concrete proposal of what it might be.

2.3 The dynamics of variational learning

We now turn to the computational properties of the variational model in (22).

2.3.1 *Asymptotic behaviors*

In any competition process, some measure of fitness is required. Adapting the formulation of Bush & Mosteller (1958), we may offer the following definition:

(23) The *penalty probability* of grammar G_i in a linguistic environment E is

$$c_i = \Pr(G_i \not\mapsto s \mid s \in E)$$

The penalty probability c_i represents the probability that a grammar G_i fails to analyze an incoming sentence and gets punished as a result. In other words, c_i is the percentage of sentences in the environment with which the grammar G_i is incompatible. Notice that penalty probability is a fixed property of a grammar relative to a fixed linguistic environment E, from which input sentences are drawn.

For example, consider a Germanic V2 environment, where the main verb is situated in the second constituent position. A V2 grammar, of course, has the penalty probability of 0.[10] An English-type SVO grammar, although not compatible with all V2 sentences, is nevertheless compatible with a certain proportion of them. According to a corpus analysis cited in Lightfoot (1997), about 70% of matrix sentences in modern V2 languages have the surface order of SVO: an SVO grammar therefore has a penalty probability of 30% in a V2 environment. Since the grammars in the delimited UG space are fixed—it is only their weights that change during learning—their fitness values defined as penalty probabilities are also fixed if the linguistic environment is, by assumption, fixed.

It is crucial to realize that penalty probability is an *extensionally* defined property of grammars. It is a notion used, by the linguist, in the formal analysis of the learning model. It is not a component of the learning process. For example, the learner needs not and does not keep track of frequency information about sentence patterns, and does not explicitly compute the penalty probabilities of the competing grammars. Nor is penalty probability represented or accessed in during learning, as the model in (22) makes clear.

[10] For expository ease we will keep to the fitness measure of whole grammars in the present discussion. In section 2.4 we will place the model in a more realistic P&P grammar space, and discuss the desirable consequences in the reduction of computational cost.

The asymptotic properties of the L_{R-P} model have been extensively studied in both mathematical psychology (Norman 1972) and machine learning (Narendra & Thathachar 1989, Barto & Sutton 1998). For simplicity but without loss of generality, suppose that there are two grammars in the population, G_1 and G_2, and that they are associated with penalty probabilities of c_1 and c_2 respectively. If the learning rate γ is sufficiently small, i.e. the learner does not alter his 'confidence' in grammars too radically, one can show (see Narendra and Thathachar 1989: 162–5) that the asymptotic distributions of $p_1(t)$ and $p_2(t)$ will be essentially normal and can be approximated as follows:

(24) Theorem:

$$\lim_{t\to\infty} p_1(t) = \frac{c_2}{c_1 + c_2}$$

$$\lim_{t\to\infty} p_2(t) = \frac{c_1}{c_1 + c_2}$$

(24) shows that in the general case, grammars more compatible with the input data are better represented in the population than those less compatible with the input data as the result of learning.

2.3.2 *Stable multiple grammars*

Recall from section 2.1.1 that realistic linguistic environments are usually heterogeneous, and the actual linguistic data cannot be attributed to a single idealized 'grammar'. This inherent variability poses a significant challenge for the robustness of the triggering model.

How does the variational model fare in realistic environments that are inherently variable? Observe that non-homogeneous linguistic expressions can be viewed as a probabilistic combination of expressions generated by multiple grammars. From a learning perspective, a non-homogeneous environment induces a population of grammars none of which is 100% compatible with the input data. The theorem in (24) shows that the weights of two

(or more, in the general case) grammars reach a stable equilibrium when learning stops. Therefore, the variability of a speaker's linguistic competence can be viewed as a probabilistic combination of multiple grammars. We note in passing that this interpretation is similar to the concept of 'variable rules' (Labov 1969, Sankoff 1978), and may offer a way to integrate generative linguists' idealized grammars with the study of language variation and use in linguistic performance. In Chapter 5, we extend the acquisition model to language change. We show that a combination of grammars as the result of acquisition, while stable in a single (synchronic) generation of learners, may not be diachronically stable. We will derive certain conditions under which one grammar will inevitably replace another in a number of generations, much like the process of natural selection. This formalizes historical linguists' intuition of grammar competition as a mechanism for language change.

Consider the special case of an idealized environment in which all linguistic expressions are generated by an input grammar G_1. By definition, G_1 has a penalty probability of 0, while all other grammars in the population have positive penalty probabilities. It is easy to see from (24) that the p_1 converges to 1, with the competing grammars eliminated. Thus, the variational model meets the traditional learnability condition.

Empirically, one of the most important features of the variational model is its ability to make quantitative predictions about language development via the calculation of the expected change in the weights of the competing grammars. Again, consider two grammars, target G_1 and the competitor G_2, with $c_1 = 0$ and $c_2 > 0$. At any time, $p_1 + p_2 = 1$. With the presentation of each input sentence, the expected increase of p_1, $E[\Delta p_1]$, can be computed as follows:

(25)
$$\begin{aligned} E[\Delta p_1] = {} & p_1\gamma(1-p_1) + && \text{with Pr. } p_1,\ G_1 \text{ is chosen and } G_1 \to s \\ & p_2(1-c_2)(-\gamma)p_1 + && \text{with Pr. } p_2(1-c_2),\ G_2 \text{ is chosen and } G_2 \to s \\ & p_2c_2\gamma(1-p_1) && \text{with Pr. } p_2c_2,\ G_2 \text{ is chosen and } G_2 \not\to s \\ = {} & c_2\gamma(1-p_1) \end{aligned}$$

Although the actual rate of language development is hard to predict—it would rely on an accurate estimate of the learning parameter and the precise manner in which the learner updates grammar weights—the model does make *comparative* predictions on language development. That is, *ceteris paribus*, the rate at which a grammar is learned is determined by the penalty probability (c_2) of its competitor. By estimating penalty probabilities of grammars from CHILDES (25) allows us to make longitudinal predictions about language development that can be verified against actual findings. In Chapter 4, we do just that.

Before we go on, a disclaimer, or rather, a confession, is in order. We in fact are not committed to the L_{R-P} model *per se*: exactly how children change grammar weights in response to their success or failure, as said earlier, is almost completely unknown. What we are committed to is the *mode* of learning: coexisting hypotheses in competition and gradual selection, as schematically illustrated in (18), and elaborated throughout this book with case studies in child language. The choice of the L_{R-P} model is justified mainly because it allows the learner to converge to a stable equilibrium of grammar weights when the linguistic evidence is not homogeneous (24). This is needed to accommodate the fact of linguistic variation in adult speakers that is particularly clear in language change, as we shall see in Chapter 5. There are doubtlessly many other models with similar properties.

2.3.3 *Unambiguous evidence*

The theorem in (24) states that in the variational model, convergence to the target grammar is guaranteed if all competitor grammars have positive penalty probabilities. One way to ensure this is to assume the existence of unambiguous evidence (Fodor 1998): sentences that are compatible only with the target grammar, and not with any other grammar. While the general existence of unambiguous evidence has been questioned (Clark 1992, Clark & Roberts 1993), the present model does not require unambiguous evidence to converge in any case.

To illustrate this, consider the following example. The target of learning is a Dutch V2 grammar, which competes in a population of (prototype) grammars, where X denotes an adverb, a prepositional phrase, and other adjuncts that can freely appear at the initial position of a sentence:

(26) a. Dutch: SVO, XVSO, OVS
b. Hebrew: SVO, XVSO
c. English: SVO, XSVO
d. Irish: VSO, XVSO
e. Hixkaryana: OVS, XOVS

The grammars in (26) are followed by some of the matrix sentences word orders they can generate/analyze.[11] Observe that none of the patterns in (26a) *alone* could distinguish Dutch from the other four human grammars, as each of them is compatible with certain V2 sentences. Specifically, based on the input evidence received by a Dutch child (Hein), we found that in declarative sentences, for which the V2 constraint is relevant, 64.7% are SVO patterns, followed by XVSO patterns at 34% and only 1.3% OVS patterns.[12] Most notably, Hebrew, and Semitic in general, grammar, which allows VSO and SVO alternations (Universal 6: Greenberg 1963; see also Fassi-Fehri 1993, Shlonsky 1997), is compatible with 98.7% of V2 sentences.

Despite the lack of unambiguous evidence for the V2 grammar, as long as SVO, OVS, and XVSO patterns appear at positive frequencies, all the competing grammars in (26) will be punished. The V2 grammar, however, is never punished. The theorem in (24) thus ensures the learner's convergence to the target V2 grammar. The competition of grammars is illustrated in Fig. 2.1, based on a computer simulation.

[11] For simplicity, we assume a degree-0 learner in the sense of Lightfoot (1991), for which we can find relevant corpus statistics in the literature.

[12] Thanks to Edith Kaan for her help in this corpus study.

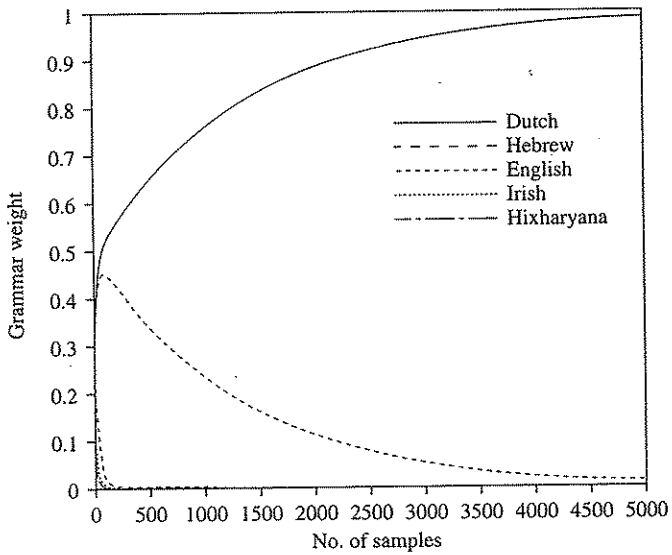

FIGURE 2.1. The convergence to the V2 grammar in the absence of unambiguous evidence

2.4 Learning grammars in a parametric space

The variational model developed in the preceding sections is entirely theory-neutral. It only requires a finite and non-arbitrary space of possible grammars, a conclusion accepted by many of today's linguists.[13] Some interesting questions arise when we situate the learning model in a realistic theory of grammar space, the P&P model.

2.4.1 *Parameter interference*

So far we have been treating competing grammars as individual entities; we have not taken into account the structure of the

[13] Different theories of UG will yield different generalizations: when situated into a theory-neutral learning model, they will—if they are not merely notational variants—make different developmental predictions. The present model can then be used as an independent procedure to evaluate linguistic theories. See Ch. 6 for a brief discussion.

grammar space. Although the convergence result in (24) for two grammars generalizes to any number of grammars, it is clear that when the number of grammars increases, the number of grammar weights that have to be stored also increases. According to some estimates (Clark 1992; cf. Kayne 2000, Baker 2001), 30–40 binary parameters are required to give a reasonable coverage of the UG space. And, if the grammars are stored as individual wholes, the learner would have to manipulate 2^{30}–2^{40} grammar weights: now *that* seems implausible.

It turns out that a parametric view of grammar variation, independently motivated by comparative theoretical linguistics, dramatically reduces the computational load of learning. Suppose that there are n binary parameters, $\alpha_1, \alpha_2, \ldots, \alpha_n$, which can specify 2^n grammars. Each parameter α_i is associated with a weight p_i, the probability of the parameter α_i being 1. The weights constitute an n-dimensional vector of real numbers between [0, 1]: $\mathrm{P} = (p_1, p_2, \ldots, p_n)$.

Now the problem of selecting a grammar becomes the problem of selecting a vector of n 0s and 1s, which can be done independently according to the parameter weights. For example, if the current value of p_i is 0.7, then the learner has a 70% chance of selecting 1 and a 30% chance of selecting 0. As the value of p_i changes, so will the probability of selecting 1 or 0. Now, given a current parameter weight vector $\mathrm{P} = (p_1, p_2, \ldots, p_n)$, the learner can non-deterministically generate a string of 0s and 1s, which is a grammar, G. Write this as $\mathrm{P} \Rightarrow G$, and the probability of $\mathrm{P} \Rightarrow G$ is the product of the parameter weights with respect to G's parameter values. P gives rise to all 2^n grammars; as P changes, the probability of $\mathrm{P} \Rightarrow G$ also changes. When P reaches the target vector, then the probability of generating non-target grammars will be infinitely small.

(27) describes how P generates a grammar to analyze an incoming sentence:

(27) For each incoming sentence s
  a. For parameter i, $i = 1, 2, \ldots, n$
    - with probability p_i, choose the value of α_i to be 1;
    - with probability $1 - p_i$, choose the value of α_i to be 0.
  b. Let G be the grammar with the parameter values chosen in (27a).
  c. Analyze s with G.
  d. Update the parameter values to $\mathbf{P}' = (p_1', p_2', \ldots, p_n')$ accordingly.

Now a problem of *parameter interference* immediately arises. Under the parametric representation of grammars, grammar selection is based on independent *parameters*. By contrast, fitness measure and thus the outcome of learning—reward or punishment—is defined on whole *grammars*. How does the learner infer, backwards, what to do with individual parameter weights, from their collective fitness as a composite grammar? In other words, what is the proper interpretation of *accordingly* in the parameter learning model (27)?

To be concrete, suppose we have two independent parameters: one determines whether the language has overt *Wh* movement (as in English but not Chinese), and the other determines whether the language has verb second (V2), generally taken to be the movement of inflected verbs to matrix Complementizer position, as in many Germanic languages. Suppose that the language to be acquired is German, which has [+Wh] and [+V2]. When the parameter combination [+Wh, –V2] is chosen, the learner is presented with a declarative sentence. Now although [+Wh] is the target value for the *Wh* parameter, the whole grammar [+Wh, –V2] is nevertheless incompatible with a V2 declarative sentence and will fail. But should the learner prevent the correct parameter value [+Wh] from being punished? If so, how? Similarly, the grammar [–Wh, +V2] will succeed at any declarative German sentence, and the wrong parameter value [–Wh], irrelevant to the input, may hitch a ride and get rewarded.

So the problem is this. The requirement of psychological plausibility forces us to cast grammar probability competition in terms of parameter probability competition. This in turns introduces the problem of parameter interference: updating independent parameter probability is made complicated by the success/failure of the composite grammar. In what follows, we will address this problem from several angles that, in combination, may yield a decent solution.

2.4.2 *Independent parameters and signatures*

To be sure, not all parameters are subject to the interference problem. Some parameters are independent of other parameters, and can be learned independently from a class of input examples that we will call *signatures*. Specifically, with respect to a parameter α, its signature refers to s_α, a class of sentences that are analyzable only if α is set to the target value. Furthermore, if the input sentence does not belong to s_α, the value of α is not material to the analyzability of that sentence.

In the variational model, unlike the cue-based learning model to be reviewed a little later, the signature–parameter association need not be specified a priori, and neither does the learner actively search for signature in the input. Rather, signatures are interpreted as input whose cumulative effect leads to correct setting of parameters. Specifically, both values of a parameter are available to the child at the outset. The non-target value, however, is penalized upon the presentation of 'signatures', which, by definition, are only compatible with the target value. Hence, the non-target value has a positive penalty probability, and will be eliminated after a sufficient number of signatures have been encountered.

The existence of signatures for independent parameters is useful in two important ways. On the one hand, it radically reduces the problem of parameter interferences. For every parameter that is independent, the learning space is in effect cut by half; we will clarify this claim shortly, in section 2.4.4.[14] On the

[14] This also suggests that when proposing syntactic parameters, we should have the problem of acquisition in mind. When possible, parameters that can be independently learned better serve the goal of explanatory adequacy in reducing the cognitive load of child language acquisition.

other hand, parameters with signatures lead to longitudinal predictions that can be directly related to corpus statistics. For two such parameters, we can estimate the frequencies of their respective signature, and predict, on the basis of (25), that the parameter with more abundant signatures be learned sooner than the other. In Chapter 4, we will see the acquisition of several independent parameters that can be developmentally tracked this way.

So what *are* these independent parameters? Of the better-established parameters, a few are obviously independent. The *Wh* movement parameter is a straightforward example. *Wh* words move in English questions, but not in Chinese questions, and *Wh* questions will serve to unambiguously determine the target values of this parameter, regardless of the values of other parameters. For non-*Wh* sentences, the *Wh* parameter obviously has no effect.

Another independent parameter is the verb raising parameter that determines whether a finite verb raises to Tense: French sets this parameter to 1, and English, 0 (Emonds 1978, Pollock 1989). The 1 value for this parameter is associated with signature such as (28), where finite verbs precede negation/adverb:[15]

(28) a. Jean ne mange pas de fromage.
Jean *ne* eats no of cheese.
'John does not eat cheese.'
b. Jean mange souvent du fromage.
Jean eats often of cheese.
'John often eats cheese.'

Yet another independent parameter is the obligatory subject parameter, for which the positive value (e.g. English) is associated with the use of pure expletives such as *there* in sentences like *There is a train in the house.*

[15] Although it is possible that the verb does not stop at Tense but raises further to higher nodes (as in verb-second environments), the principle of the Head Movement Constraint (Travis 1984), or more generally economy conditions (Chomsky 1995b), would prohibit such raising to skip the intermediate Tense node. Therefore, finite verbs followed by negation or adverbs in a language indicate that the verb must raise at least to Tense.

What about the parameters are not independent, whose values can not be directly determined by any particular type of input data? In section 2.4.3 we review two models that untangle parameter interference by endowing the learner with additional resources. We then propose, in section 2.4.4, a far simpler model and study its formal sufficiency. Our discussion is somewhat technical; the disinterested reader can go straight to section 2.5. A fuller treatment of the mathematical and computational issues can be found in Yang (in press).

2.4.3 *Interference avoidance models*

One approach is to give the learner the ability to tease out the relevance of parameters with respect of an input sentence. Fodor's (1998) Structural Trigger Learner (STL) takes this approach. The STL has access to a special parser that can detect whether an input sentence is parametrically ambiguous. If so, the present parameter values are left unchanged; parameters are set only when the input is completely unambiguous. The STL thus aims to avoid the local maxima problem, caused by parametric inference, in Gibson & Wexler's triggering model.[16]

The other approach was proposed by Dresher & Kaye (1990) and Dresher (1999); see Lightfoot (1999) for an extension to the acquisition of syntax. They note that the parameters in metrical stress can be associated with a corresponding set of *cues*, input data that can unambiguously determine the values of the parameters in a language. Dresher & Kaye (1990) propose that for each parameter, the learner is innately endowed with the knowledge of the cue associated with that parameter. In addition, each parameter has a *default* value, which is innately specified as well. Upon the presentation of a cue, the learner sets the value for the corresponding parameter. Crucially, cues are *ordered*. That is, the cue

[16] Tesar & Smolensky Constraint Demotion model (2000) is similar. For them, a pair of violable constraints is (re)ordered only when their relative ranking can be unambiguously determined from an input datum; the detection of ambiguity involves examining other candidate rankings.

for a parameter may not be usable if another parameter has not been set. This leads to a particular sequence of parameter setting, which must be innately specified. Suppose the parameter sequence is $\alpha_1, \alpha_2, \ldots \alpha_n$, associated with cues $s_1, s_2, \ldots, s_n$, respectively. (29) schematically shows the mechanisms of the cue-based learner:

(29) a. Initialize $\alpha_1, \alpha_2, \ldots, \alpha_n$ with their respective default values.
b. For $i = 1, 2, \ldots, n$
- Set α_i upon seeing s_i.
- Leave the set parameters $\alpha_1, \ldots, \alpha_{n-1}$ alone.
- Reset $\alpha_{i+1}, \ldots, \alpha_n$ to respective default values.

In the present context, we do not discuss the formal sufficiency of the STL and the cue-based models.[17] The STL model seems to introduce computational cost that is too high to be realistic: the learner faces a very large degree of structural ambiguity that must be disentangled (Sakas & Fodor 2001). The cue-based model would only work if *all* parameters are associated with cues and default values, and the order in which parameters are set must be identified as well. While this has been deductively worked out for about a dozen parameters in metrical stress (Dresher 1999), whether the same is true for a non-trivial space of syntactic parameters remains to be seen.

Both models run into problems with the developmental compatibility condition, detrimental to all transformational learning models: they cannot capture the variation in and the gradualness of language development. The STL model may maintain that before a parameter is conclusively set, both parameter values are available, to which variation in child language are be attributed. However, when a parameter *is* set, it is set in an all-or-none fashion, which then incorrectly predicts abrupt changes in child language.

The cue-based model is completely deterministic. At any time,

[17] Both have problems: see Bertolo et al. (1997) for a formal discussion; see also Church (1992) for general comments on the cue-based model, and Gillis et al. (1995) for a computer simulation.

a parameter is associated with a unique parameter value—correct or incorrect, but not both—and hence no variation in child language can be accounted for. In addition, the unset parameters are reset to default values every time a parameter is set. This predicts radical and abrupt reorganization of child language: incorrectly, as reviewed earlier. Finally, the cue-based model entails that learners of all languages will follow an identical learning path, the order in which parameters are set: we have not been able to evaluate this claim.

2.4.4 *Naive parameter learning*

In what follows, we will pursue an approach that sticks to the strategy of assuming a 'dumb' learner.[18] Consider the algorithm in (30), a *Naive Parameter Learner* (NPL):

(30) Naive Parameter Learning (NPL)
a. Reward *all* the parameter values if the composite grammar succeeds.
b. Punish *all* the parameter values if the composite grammar fails.

The NPL model may reward wrong parameter values as hitchhikers, and punish correct parameter values as accomplices. The hope is that, in the long run, the correct parameter values will prevail.

To see how (30) works, consider again the learning of the two parameters [Wh] and [V2] in a German environment. The combinations of the two parameters give four grammars, of which we can explicitly measure the fitness values (penalty probabilities). Based on the CHILDES corpus, we estimate that about 30% of all sentences children hear are Wh questions,[19] which are only compatible with the [+Wh] value. Of the remaining declarative sentences, about 49% are SVO sentences that are consistent with the [–V2] value. The other 21% are VS sentences with a topic

[18] For useful discussions I would like to thank Sam Gutmann, Julie Legate, and in particular Morgan Sonderegger for presenting our joint work here.

[19] This figure is based on English data: we are taking the liberty to extrapolate it to our (hypothetical) German simulation.

in [Spec,CP], which are only compatible with the [+V2] value. We then have the penalty probabilities shown in Table 2.1.

Fig. 2.2 shows the changes of the two parameter values over time. We see that the two parameters, which fluctuated in earlier stages of learning—the target values were punished and the non-target values were rewarded—converged correctly to [1, 1] in the end.

It is not difficult to prove that for parameters with signatures, the NPL will converge on the target value, using the Martingale methods in Yang & Gutmann (1999); see Yang (in press) for

TABLE 2.1. The penalty probabilities of four grammars composed of two parameters

	[+Wh]	[–Wh]
[+V2]	0	0.3
[–V2]	0.21	0.51

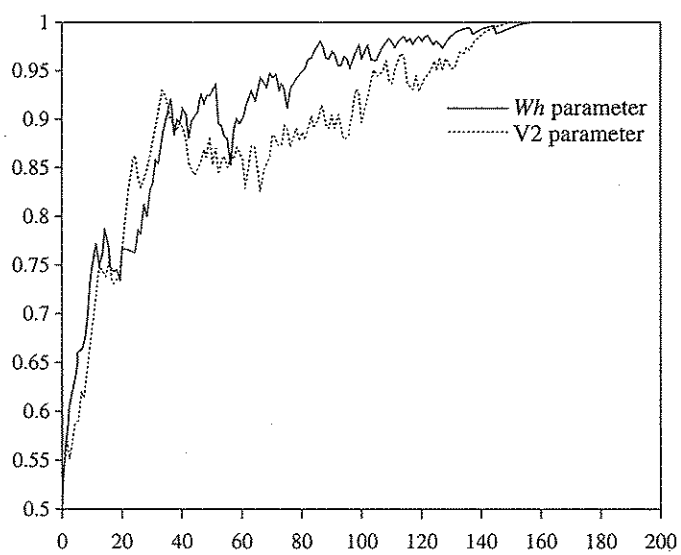

FIGURE 2.2. The independent learning of two parameters, Wh and V2

details. We now turn to the more difficult issue of learning parameters that are subject to the interference problem.

Fitness distribution

In what follows, we will suggest that (some variant) of the NPL may be a plausible model of learning that distangles the interference effects from parameter interaction.

First, our conclusion is based on results from computer simulation. This is not the preferred move, for the obvious reason that one cannot simulate all possibilities that may arise in parameter learning. Analytical results—proofs—are much better, but so far they have been elusive.

Second, as far as feasible, we will study the behavior of the model in an *actual* learning environment. As the example of the *Wh* and V2 learning (Fig. 2.2) shows, the relative fitness values of the four composite grammars will determine the outcome of parameter learning. In that example, if the three competitors have high penalty probabilities, intuition tells us that the two parameters rise to target values quickly.[20] So the actual behavior of the model can be understood only if we have a good handle on the fitness distribution of actual grammars.

This is a departure from the traditional linguistic learnability study, and we believe it is a necessary one. Learnability models, in general, do not consider convergence in relation to the actual (statistical) distribution of the learning data.[21] Rather, learning is studied 'in the limit' (Gold 1967), with the assumption that learning can take an arbitrary amount of data as long as it converges on the correct grammar in the end: hence, no sample complexity considerations. However, it is clear that learning data is not infinite. In Chapter 4 we show that it is possible to establish bounds on the amount of linguistic data needed for actual acquisition: if

[20] Although intuition fades rapidly as more and more parameters combine and interact.

[21] A notable exception is Berwick & Niyogi's (1996) elegant Markov model of triggering, where the expected amount of evidence required for convergence can be precisely worked out.

the learning data required by a model greatly exceed such bounds, then such a model will fail the formal sufficiency condition.

Sample complexity, even if it *is* formally studied, means very little unless placed in an actual context. For example, suppose one has found models that require exactly n or n^2 specific kinds of input sentences to set n parameters. The sample complexity of this model is very small: a (low) polynomial function of the problem size. But to claim this is an efficient model, one must show that these n^2 sentences are in fact *attested* with robust frequencies in the actual input: a model whose theoretical convergence relies on twenty levels of embedded clauses with parasitic gaps is hopeless in reality.

In a similar vein, a model that fails under some hypothetical conditions may not be doomed either: it is possible that such cases never arise in actual learning environments. For example, computer simulation shows that the NPL model does not converge onto the target parameter values in a reasonable amount of time if all of the $2^n - 1$ composite grammars have the penalty probability of 0.1: that is, all non-target grammars are equally good, compatible with 90% of input data. But this curious (and disastrous) scenario does not occur in reality.

It is very difficult to know what actual penalty probability distributions are like. To do so, one would have to consider all, at least a large portion, of the 2^n grammars. For each grammar, which is a parameter value vector, one needs to find a corresponding existing language, take a large sample of sentences from it, and then analyze the sample with all the other $2^n - 1$ competitors. It is obvious that each of these steps poses enormous practical problems for large numbers of n. Our experience working with corpora (Chapter 4) suggests that there are relatively few competing grammars with low penalty probabilities, i.e. very close to the target grammar, whereas the vast majority of them are bad. The example in (26), the V2 grammar in competition with four other grammars, is a case in point. This assumption seems compatible with the fact that most (but not all) parameters are acquired fairly early, which would not be possible if the relative compatibilities among grammars were very high.

Furthermore, we believe that it is reasonable to assume that the badness of a grammar is in general correlated with how 'far away' it is from the target grammar, where distance can be measured by how many parameter values they differ: the Hamming distance. In particular, we assume that as grammars get further and further away, their fitness values deteriorate rapidly. It is true that the change of *some* parameter may induce radical changes on the overall grammar obtained, e.g. [±Wh], scrambling (though some of these parameters may be independent, and thus free of parameter interference). Hence, what we assume is only a statistical tendency: it is possible that a grammar closer to the target (in terms of the Hamming distance) is worse than one that is further away, but it is unlikely.

Specifically, we assume that the penalty probabilities of the competing grammars follow a standard Gaussian distribution:

(31) $c(x) = 1 - e^{-\frac{x^2}{2\sigma^2}}$, where $\sigma = 1/3$

To choose penalty probabilities, we first divide the interval (0, 1) into n equal segments, where n is the number of parameters. A grammar G_h with Hamming distance h is expected to fall in the hth interval. However, to simulate the effect that grammars further from the target are generally (but not always) worse than the closer ones, we assume that G_h falls in the hth region with probability s, in the $h \pm 1$st regions with probability s^2, in the $h \pm 2$nd regions with probability s^3, etc. This is our assumption of *exponential decay* of grammar fitness with respect to its Hamming distance. Thus, a grammar farther away can be still be compatible with many sentences from the target grammar, but the likelihood of it being so vanishes very quickly. Similarly, a grammar that differs from the target by few paprameters can also be fairly bad. But overall, further away grammars are on average worse than those that are closer to the target.

To verify our assumptions of penalty probability distributions, we consider a very small case, for $n = 3$ with three parameters, in Gibson & Wexler (1994): Spec-Head, Comp-Head, and V2. And

even here we will make simplified assumptions; see Appendix A for details. First, we only consider the matrix clauses, as in Gibson & Wexler (1994). Second, some essential distributional statistics are based on English and Germanic languages, and then extrapolated (not unreasonably, we believe) to other grammars. Averaging over the pairwise penalty probabilities of eight grammars, we have:

(32) a. The average penalty probability for grammars one parameter away is 0.571312.
b. The average penalty probability for grammars two parameters away is 0.687908.
c. The average penalty probability for grammars three parameters away is 0.727075.

This is clearly consistent with our assumption about fitness distribution. Penalty probability in general correlates with the Hamming distance from the target. The pairwise penalty probabilities (Table 2.3 in Appendix A) are also consistent with our assumption of distance-related exponential decay.

2.4.5 *Learning rates and random walks*

If one runs the NPL on the distribution of penalty probabilities as in (31), a number of problems arise, all having to do with the choice of the learning parameter, γ, which controls the rapidity with which the learner adjusts the parameters. First, if γ is too small—the learner modifies parameter weights very slightly upon success/failure—the learner takes an incredibly long time to converge. And second, if γ is too big, the learner will modify the parameter weights very abruptly, resulting in a 'jumpy' learning curve, not so unlike the original triggering model rejected on the ground of developmental incompatibility (section 2.1.2).

It is not hard to understand why this may be the case. Consider the current parameter weight vector $\mathrm{P} = (p_1, p_2, \ldots, p_n)$, and the target values are T, which is an n-ary vector of 0s and 1s. When P is far from T, e.g. $\mathrm{P} = (0.5, 0.5, \ldots, 0.5)$, the learner has no idea what T may be. As P gets closer to T, the learner will be able to

analyze incoming sentences more often. Thus, the learner may have increasingly higher confidence in P, which now works better and better. It then seems reasonable to assume that the learner ought to be more conservative when P is far from the target, but more assured when P gets close.

There are a number of ways of implementing this intuition. One may assume that the gradual increase in γ is a matter of biological maturation. There are also many algorithms in computer science and machine learning that formally—and computationally expensively—modify the learning rate with respect to the confidence interval. But these approaches will alter the mathematical properties of the L_{R-P} model (22), which requires a fixed learning rate. Furthermore, they deviate from the guidelines of psychological plausibility and explanatory continuity that acquisition models are advised to follow (Chapter 1).

An alternative is suggested by Morgan Sonderegger (personal communication). It is based on two observations. First, note that having a high γ is equivalent to having a fixed γ and *using it* often. Second, the overall goodness of P can be related to how often P successfully analyzes incoming sentences. This leads to a very simple measure of how close P is to the target, by introducing a small batch counter b, which is initialized to 0, and a batch bound B, a small positive integer (usually between 2 and 5, in practice). Formally,

(33) The Naive Parameter Learner with Batch (NPL+B)
a. For an input sentence s, select a grammar G based on P following the procedure in (27)
b. • If $G \rightarrow s$, then $b = b + 1$.
• If $G \nrightarrow s$, then $b = b - 1$.
c. • If $b = B$, reward G and reset $b = 0$.
• If $b = -B$, punish G and reset $b = 0$.
d. Go to (33a).

Note that the use of 'batch' in NPL+B (33) is very different from the standard one. Usually, 'batch' refers to a memory that stores a number of data points before processing them. In NPL+B, b is

simply a counter that tracks the success or failure of sentence analysis, without recording what sentences have been presented or what grammars selected. The cost of additional memory load is trivial.

Yet the effect of this batch is precisely what we wanted: it slows down the learning rate when P is bad, and speeds it up when P gets better. To see this, consider that P is very close to T. Now almost every sentence is compatible with the grammars given by P, because most of the non-target grammars now have a very low probability of being selected. Then, almost every B sentences will push the batch counter b to its bound (B). Weights will be updated very frequently, driving P to T ever more rapidly. By contrast, if P is quite far from T, then it generally takes a longer time for b to reach its bound—reward and punishment are then less frequent, and thus slow down learning.

This batch process can be understood precisely by considering the problem of the Gambler's Ruin. A gambler has n dollars to start the game. Every bet he makes, there is a probability p of making a dollar, and a probability $q = 1 - p$ of losing a dollar. The gambler *wins* if he ends up with $2n$ dollars, and is ruined if he is down to 0. Since every gamble is independent of all others, the gambler's fortune takes a random walk. It is not difficult to show—the interested reader may consult any textbook on stochastic processes—that the probability of the gambler winning (i.e. getting $2n$ dollars), w, is:

$$(34) \quad w = \frac{(q/p)^n - 1}{(q/p)^{2n} - 1}$$

Our batch counter b does exactly the same thing. It gains 1 when P yields a succesful grammar, and loses 1 when P yields a failing grammar. b wins if it reaches B, and loses if it reaches $-B$. Let p be the probability of P yielding a successful grammar.[22]

[22] Precisely, $p = \sum_i \Pr(P \Rightarrow G_i)(1 - c_i)$, where G_i is a grammar that can be generated by P (there are 2^n such grammars), where n is the number of parameters, $\Pr(P \Rightarrow G_i)$ is the probability that P generates the grammar G_i (see (27)), and c_i is the penalty probability of G_i.

Then w_B, the probability of b reaching the batch bound B is

$$(35) \quad w(B, p) = \frac{(q/p)^B - 1}{(q/p)^{2B} - 1}$$

Clearly, as p gets bigger, w_B gets larger, and as B increases, w_B gets larger still. Fig. 2.3 shows $w(B, p)$ as a function of B and p. $B = 1$ means that there is no batch: the learning parameter would be uniform throughout learning.

The assumptions of the normal distribution of grammar fitness, the exponential decay of fitness with respect to the Hamming distance, and the use of a small batch counter together give rise to a satisfactory learner, the NPL+B model.[23] A typical result from a simulation of learning a ten-parameter grammar is given in Fig. 2.4.

The learning curve is generally smooth, with no abrupt changes. And the learner converges in a reasonable amount of time. About 600,000 sentences were needed for converging on ten interacting parameters.

It must be conceded that the formal sufficiency condition of the NPL model is only tentatively established. Future research lies in two directions. First, and obviously, much more work is needed to establish whether the assumptions of Gaussian distribution and exponential decay are accurate. Second, one may (manually) determine how many parameters are in fact independent, and thus do not lead to parameter interference.[24]

The most important consequence of the NLP model, if vindicated, lies in the dramatic reduction of computational cost: the memory load reduced from storing 2^n grammar weights to n parameter weights. This makes the variational model psychologically plausible, and in turn gives a computational argument for the conception of UG as a parametric space.

[23] A copy of the NPL+B learner can be obtained from the author.

[24] If abundant, then it is good news for the STL model (Fodor 1998, Sakas & Fodor 2001). Presumably, the learner can focus on parameters that are not independent: a smaller space means smaller computational cost for the STL parser.

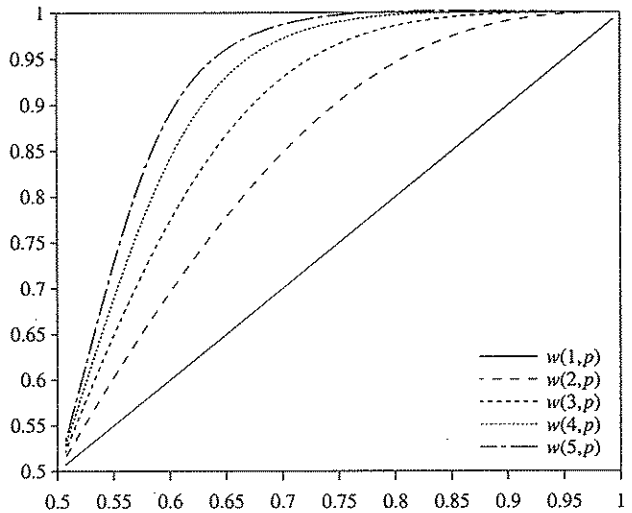

FIGURE 2.3. The probability function $w(B, p) = \frac{(q/p)^B - 1}{(q/p)^{2B} - 1}$

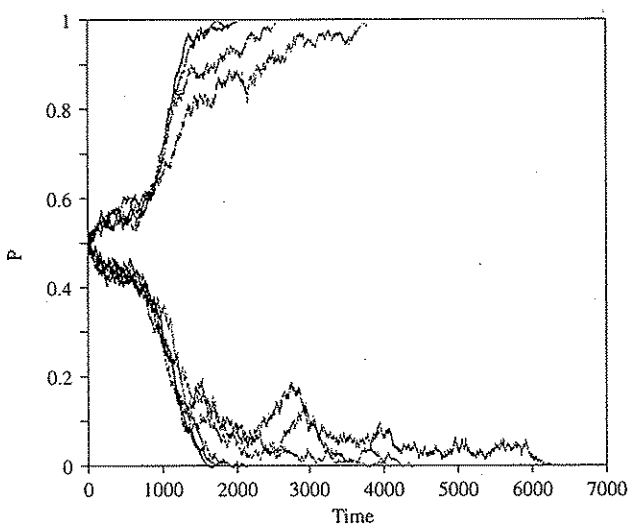

FIGURE 2.4. NPL+B learning of the grammar (1 0 0 0 0 1 0 1 1 0). B = 5, γ = 0.002, s = 5/6

Learning in a parametric space gives rise to 'hybrid' grammars. Since the successful acquisition of a grammar is accomplished only when all parameters are set correctly, children may go through an extended period of time in which some parameters are already in place while others are still fluctuating. For example, an English child may have learned that his language moves *Wh* words overtly, but has not conclusively learned that it also obligatorily uses overt subjects. Now what the child possesses are partial fragments of grammars that may not correspond to any attested adult language—something that is, say, English-like in one respect but Chinese-like in another. And it is precisely these hybrid languages that confirms the reality of grammar coexistence and competition. A number of such cases in child languages will be documented in Chapter 4.

2.5 Related approaches

The idea of language acquisition as grammar competition has occasionally surfaced in the literature, although it has never been pursued systematically or directly related to quantitative data in language development.

To the best of our knowledge, Jakobson (1941) was the first to interpret 'errors' in child phonology as possible phonological forms in non-target adult languages. This position was echoed in Stampe (1979), and seems to be accepted by at least some researchers in phonological acquisition (Macken 1995). Recent studies on infants' gradual loss of universal ability for phonetic discrimination (Kuhl et al. 1992; cf. de Boysson-Bardies 1999) seem to suggest that the variational model, in which the hypothesis space goes from 'more' to 'less' through competition, may hint at a general process that also governs the development of phonetic perception.

Since the advent of the P&P framework, some linguists have claimed that syntactic acquisition selects a grammar out of all possible human grammars (Piattelli-Palmarini 1989, Lightfoot 1991), but nothing has been formalized. That children may have simultaneous

access to multiple hypotheses has been suggested by Berwick & Weinberg (1984) and Pinker (1984), among others. The possibility of associating grammars with weights has been raised by Valian (1990), Weinberg (1990), and Bloom (1993), either for learnability considerations or to explain the gradual developmental patterns in child language. These authors, however, opted for different solutions to the problems under study.

Recently, Roeper (2000; cf. Yang 2000) has independently proposed that child language be explained as a combination of multiple grammars simultaneously available to the learner. Roeper further suggests that in the selection of competing grammars, the learner follows some principles of economy akin to those in the Minimalist Program (Chomsky 1995b): grammars with less complex structural representations are preferred.[25] Roeper gives evidence for the view of multiple grammars. For instance, English children who alternate between *I go*, using a nominative case subject, and *me go*, using a default (accusative) case, can be viewed as using two grammars with different case/agreement systems, both of which are attested in human languages.

The genetic algorithm (GA) model of Clark (1992) is most similar to the present model. The GA model represents grammars as parameter vectors, which undergo reproduction via 'crossover', i.e. parts of two parental parameter vectors are swapped and combined.[26] A mutation process is also assumed which, with some probability, randomly flips bits in the grammar vector. Candidate grammars are evaluated against input data; hence, measure of fitness is defined, which is subsequently translated into differential reproduction.

[25] The present model is presented in the most general way: all grammars are there to begin with, and input-grammar compatibility is the only criterion for rewarding/punishing grammars. It can incorporate other possibilities, including the economy condition suggested by Roeper. For instance, one can build in some appropriate prior bias in grammar evaluation—analyzability of $G \rightarrow s$ in (22)—that goes against complex grammars. However, these additional biases must be argued for empirically.

[26] This operation seems to require some empirical justification.

Both the GA model and the variational model are explicitly built on the idea of language acquisition as grammar competition; and in both models, grammars are selected for or against on the basis of their compatibility with input data. There are, however, a few important differences. One major difference lies in the evaluation of grammar fitness. In the present model, the fitness of a grammar is defined as its penalty probability, an extensional notion that is only used to described the dynamics of learning. It is not accessed by the learner, but can be measured from text corpora by the linguist. In the GA model, the learner first computes the degree of parsability for all grammars over a large sample of sentences. The parsability measures are then explicitly used to determine the differential reproduction that leads to the next generation of grammars. The computational cost associated with fitness evaluation is too large to be plausible. The variational model developed here sidesteps these problems by making use of probabilities/weights to capture the cumulative effects of discriminating linguistic evidence.

In the following chapters, we will pursue the condition of developmental compatibility and present a diverse array of evidence to support the variational model.

Appendix A: Fitness distribution in a three-parameter space

Gibson & Wexler (1994: table 3) considered the variations of degree-0 sentences within three parameters: Spec-Head, Comp-Head, and V2. The strings are composed of Subject, Verb, Object, Double Objects, Auxiliary, and Adverb (which broadly refers to adjuncts or topics that quite freely appear in the initial position of a sentence). For simplicity, we do not consider double objects. The grammars and the patterns they can generate are given in Table 2.2.

A principled way to estimate the probability of a string $w_1^n = w_1, w_2 \ldots w_n$ is to compute its joint probability by the use of the Chain Rule:

TABLE 2.2. A space of three parameters, or eight grammars, and the string patterns they allow

Language	Spec-Head	Comp-Head	V2	degree-0 sentences
VOS–V2	1	1	0	VS VOS AVS AVOS XVS XVOS XAVOS
VOS+V2	1	1	1	SV SVO OVS SV SAVO OAVS XVS XVOS XAVS XAVOS
SVO–V2	0	1	0	SV SVO SAV SAVO XSV XSVO XSAV XSAVO
SVO+V2	0	1	1	SV SVO OVS SAV SAVO OASV XVS XVSO XASV XASVO
OVS–V2	1	0	0	VS OVS VAS OVAS XVS XOVS XVAS XOVAS
OVS+V2	1	0	1	SV OVS SVO SAV SAOV OAVS XVS XVOS XAVS XAOVS
SOV–V2	0	0	0	SV SOV SVA SOVA XSOV XSVA XSOVA
SOV+V2	0	0	1	SV SVO OVS SAV SAOV OASV XVS XVSO XASV XASOV

$$p(w_1^n) = p(w_1)p(w_2|w_1)p(w_3|w_1^2)\ldots p(w_n|w_1^{n-1}) = \prod_{k=1}^{n}(w_k|w_1^{k-1})$$

where the conditional probabilities can be estimated individually. For example, if w_1 = S, w_2 = V, and w_3 = O, then $p(SVO) = p(S)p(V|S)p(O|SV)$. It is easy to estimate $p(S)$: $p(S) = 1$ for obligatory subject languages, and $p(S) < 1$ for subject drop languages. Presumably $p(V|S) = 1$: every sentence has a verb (including auxiliary verbs). And $p(O|SV)$ is simply the frequency of transitive verb uses. When the n gets large, the conditional probabilities get complicated, as substrings of $w_1 \ldots w_n$ are dependent. However, even with a very modest n, say, 10, one can get a fairly comprehensive coverage of sentential patterns (Kohl 1999). And again there is independence to be exploited; for example, verb-to-tense raising parameter is conditioned only upon the presence of a negation or adverb, and nothing else.

The crucial assumption we make is that there are similarities in the distributions of w_is across languages, no matter how these languages put them together. It does not seem unreasonable to assume, say, that the frequencies of transitive verbs are more or less uniform across languages, because transitive verbs are used in certain life contexts, which perhaps do not vary greatly across languages. Practically, such assumptions are necessary if there is any hope of estimating the distribution of sentences in many grammars, without reliable parsers or comprehensive corpora. Furthermore, some grammars, i.e. parameter settings, may not be attested in the world.

Given these assumptions, let us see how we may estimate the string distributions for eight grammars in Table 2.2, extrapolating from the grammars for which we do have some statistical results. For the English grammar (SVO–V2), we estimate, using sources like the CHILDES corpus, that about 10% of declarative sentences have an sentence-initial XP; thus 90% of the probability mass will be distributed among SV, SVO, SAV, SAVO. Roughly 50% of all sentences contain an auxiliary, and 50% of verbs are transitives. Assuming that the selection of Auxiliary and Verb is independent, and that the selection of the XP adjunct is independent of the rest of the sentence. We then obtain:

(36) a. P(SV) = P(SVO) = P(SAV) = P(SAVO) = 9/40
b. P(XSV) = P(XSVO) = P(SAV) = P(XSAVO) = 1/40

(36) will be carried over to the other three non-V2 grammars, and assigned to their respective canonical word orders.

For the four V2 grammars, we assume that (36) will carry over to the canonical patterns due to the Spec-Head and Comp-Head parameters. In addition, we must consider the effect of V2: raising S, O, or X to the sentence-initial position. It is known from (Lightfoot 1997: 265) as well as from our own analysis of a Dutch adult-to-child corpus, that in V2 languages, S occupies the initial position 70% of time, X, 28%, and O, 2%. These probability masses (0.7, 0.28, and 0.02) will be distributed among the canonical patterns.

Putting these together, we may compute the penalty probability c_{ij} of grammar G_i relative to grammar G_j:

$$c_{ij} = \sum_{G_j \to s} P(s|G_i \not\to s)$$

The pairwise c_{ij}s are given in Table 2.3.

TABLE 2.3. Relative penalty probabilities of the eight grammars

C_{ij}	G_{110}	G_{111}	G_{100}	G_{101}	G_{010}	G_{011}	G_{000}	G_{001}
G_{110}	–	0.790	1.000	0.930	0.750	0.860	0.800	0.930
G_{111}	0.900	–	0.100	0.220	0.750	0.245	0.625	0.395
G_{100}	0.999	0.300	–	0.300	1.000	0.475	0.600	0.475
G_{101}	0.966	0.220	0.100	–	0.750	0.395	0.625	0.245
G_{010}	0.742	0.920	1.000	0.920	–	0.920	0.800	0.920
G_{011}	0.933	0.245	0.325	0.395	0.750	–	0.625	0.220
G_{000}	0.999	0.825	0.750	0.825	1.000	0.825	–	0.825
G_{001}	0.967	0.395	0.325	0.245	0.750	0.200	0.625	–

Currently, we are extending these methods to grammars in a larger parametric space, based on the work of Kohl (1999).

3

Rules over Words

Fuck these irregular verbs.

Quang Phuc Dong, *English Sentences without Overt Grammatical Subject* (1971), p. 4

The acquisition of English past tense has generated much interest and controversy in cognitive science, often pitched as a clash between generative linguistics and connectionism (Rumelhart & McClelland 1986), or even between rationalism and empiricism (Pinker 1999). This is irregular: the problem of past tense, particularly in English, notorious for its impoverished phonology, is a marginal problem in linguistics, and placing it at the center of attention does no justice to the intricacy of the study of language; see e.g. Halle (2000), Yang (2000), and Embick & Marantz (in press).

Yet this is not to say the problem of English past tense is trivial or uninteresting. As we shall see, despite the enthusiasm and efforts on both sides of the debate, there remain many important patterns in the published sources still unknown and unexplained. We show that the variational learning model, instantiated here as competition among phonological rules (rather than grammars/parameters, as in the case of syntactic acquisition), provides a new understanding of how phonology is organized and learned.

3.1 Background

Our problem primarily concerns three systematic patterns in children's acquisition of past tense. First, it has been known since