Knowledge and Learning in Natural Language

by

Charles D. Yang

B.S., Case Western Reserve University (1994)
S.M., Massachusetts Institute of Technology (1997)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2000

[February 2001]

Author_____
Department of Electrical Engineering and Computer Science
August 2000

Certified by_____
Robert C. Berwick
Thesis Supervisor

Certified by_____
Noam Chomsky
Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Committee on Graduate Students

Knowledge and Learning in Natural Language

by

Charles D. Yang

Submitted to the
Department of Electrical Engineering and Computer Science

August 2000

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

Abstract

The present dissertation is a study of language development in children. From a biological perspective, the development of language, as the development of any other organic systems, is an interaction between internal and external factors; specifically, between the child's internal knowledge of linguistic structures and the external linguistic experience he receives. Drawing insights from the study of biological evolution, we put forth a quantitative model of language acquisition that make this interaction precise, by embedding a theory of knowledge, the Universal Grammar, into a theory of learning from experience. In particular, we advance the idea that language acquisition should be modeled as a population of grammatical hypotheses, competing to match the external linguistic experiences, much like in a natural selection process. We present evidence — conceptual, mathematical, and empirical, and from a number of independent areas of linguistic research, including the acquisition of syntax and morphophology, and historical language change — to demonstrate the model's correctness and utility.

Thesis Supervisor: Robert C. Berwick
Title: Professor of Computer Science and Brain and Cognitive Science

Thesis Supervisor: Noam Chomsky
Title: Institute Professor of Linguistics and Philosophy

To my parents and gramma.

"My sons, dig in the vineyard," were the last words of the old man in the fable: and, though the sons found no treasure, they made their fortune by the grapes.

— T. H. Huxley

Acknowledgments

The heresies advocated in these pages, if of any value at all, result from digging for treasures in a vineyard but harvesting grapes instead. Or, in less romantic terms, the present thesis started out a pure accident.

In the spring of 1998, I took an excellent course on evolutionary biology with two of the best evolutionary biologists around: Dick Lewontin and Stephen Jay Gould. There were plenty of treasures but I managed to find some grapes as well. On a field trip to the American Museum of Natural History, while waiting impatiently for SJG for a guided tour, I started on the class reading material, Dick Lewontin's 1983 paper *Organism as Subject and Object of Evolution*, which points out a conceptual feature of evolutionary process that cannot be exaggerated. Evolutionary change is a *variational* process. That is, it is not that individuals undergo direct change themselves; rather, it is the distribution of different individuals that changes under evolutionary forces. At that very moment, it occurred to me that language acquisition might be understood in a similar fashion: the distribution of Universal Grammar adaptively changes to the linguistic evidence presented to the child learner. The thesis explores the consequences of this idea.

Of course, knowing where to dig was only a start. In cultivating grapes I have received enormous help from my teachers and committee members. Noam Chomsky and Bob Berwick, my co-supervisors, are jointly responsible for every step in and every aspect of my intellectual development such that words of gratitude are hardly needed for they could not be adequate. Many thanks to Tom Roeper, who kindly took me under his wings and whose energy is only matched by his immense knowledge of child language. Many thanks also to Patrick Winston, for his good scientific sense and much needed support and advice during the final stage of my graduate school.

The list of supervisors does not stop with the committee: the unofficial members of the committee have been equally instrumental, and I thank them all wholeheartedly. Comrades John Frampton and Sam Gutmann, fellow members of Team Rocket Science, insisted on getting the math right and always made sure my thinking is without inconsistencies; I look forward to many more meetings at Toscanini's. Morris Halle taught me everything I know about phonology and devoted an embarrassing amount of time to this work, particularly the chapter on irregular verbs, which is dedicated to him. And, of course, Julie Legate, my best friend and critic and also member of the Team, who read everything I've ever written and contributed more than anyone else to this thesis in numerous ways.

In addition, I am indebted to many teachers, colleagues, and friends, too numerous to list here, for suggestions and comments on this thesis and other projects, and for being an important part of my graduate education. I especially want to thank Eric Grimson for clearing many hurdles during my time at MIT. The financial support for my education came from a National Science Foundation fellowship, supplemented by various grants from Bob Berwick, no string attached. Gratitude also goes to the NEC Research Institute, Sun Microsystems Research Labs, and Microsoft Research Institute, where I held summer internships, which not only provided additional income but also helped shape my perspectives on many things.

The boring life of a graduate student has been made less so by friends in the MIT AI Lab, a unique and remarkable place where free thinking, free research, free lunch, and free fun are institutionally encouraged and pursued. I wish graduate students in all disciplines at all places could be equally free. Thanks to Andrew Berry for tips on how not to make it in academia, Jeff Brennan for coffee and trash talk, Adriane Desmond for two superb intellectual biographies, Dave Evans for soccer and beer, Shigeru Miyamoto for Zelda, Carolyn Smallwood for scrabble and Jack, and the past residents of the Church Street Inn for all the above. Special thanks to Frank Wolff, who gave me a copy of *Language and Mind* when I was a college senior.

Julie has proved that doing linguistics was the best decision I've ever made: this thesis certainly would not be possible without her, but neither would much much more. In fact, life without her is unthinkable. Russell has given new meaning to what this is all about: I only wish he knew it as well. I thank our families for their support and encouragement. My

parents and my gramma are the best teachers I've ever had: to them I dedicate this thesis.

Now, on to the vineyard.

Contents

1 The Study of Language and Language Acquisition **10**

1.1 The Naturalistic Approach to Language 10

1.2 The Structure of Language Acquisition 13

1.2.1 Formal Sufficiency 14

1.2.2 Developmental Compatibility 15

1.2.3 Explanatory Continuity 17

1.3 A Roadmap 19

2 A Variational Model of Language Acquisition **21**

2.1 Against Transformational Learning 22

2.1.1 Formal Insufficiency of the Triggering Model 24

2.1.2 Developmental Incompatibility of the Triggering Model 26

2.1.3 Imperfection in Child Language? 28

2.2 The Variational Approach to Language Acquisition 30

2.2.1 The Dynamics of Darwinian Evolution 30

2.2.2 Language Acquisition as Grammar Competition 31

2.3 The Dynamics of Variational Learning 35

2.3.1 Asymptotic Behaviors 35

2.3.2 Stable Multiple Grammars 37

2.3.3 Unambiguous evidence 39

2.4 Learning Grammars in a Parametric Space 41

2.4.1 Parameters Make Learning Efficient 41

2.4.2 Parameter Expression and Cues 44
2.5 Related Approaches 45

3 Competing Grammars in Child Syntax **48**
3.1 The Time Courses of Three Parameters 49
3.1.1 Verb Raising and Subject Drop: the Baselines 50
3.1.2 V1 Patterns in V2 Learners 51
3.2 A Quantitative Argument from the Poverty of Stimulus 54
3.3 The Nature of Null Subjects in Children 57
3.3.1 The Early Acquisition of Chinese and Italian Subject Drop 59
3.3.2 Why English Kids (Sometimes) Use Chinese 61
3.4 Summary 63

4 Words, Rules, and Competitions **65**
4.1 Background 65
4.2 A Model of Rule Competition 68
4.2.1 A Simple Learning Task 69
4.2.2 Rules 70
4.2.3 Rule Competition 73
4.2.4 The Absolute and Stochastic Blocking Principles 78
4.3 Words vs. Rules in Overregularization 79
4.3.1 The Mechanics of the WR Model 80
4.3.2 The Data 82
4.3.3 Frequency Hierarchy in Verb Classes 83
4.3.4 The Free-rider effect 86
4.3.5 The Effect of Phonological Regularity: Vowel Shortening 88
4.4 Analogy, Regularity, and Rules 89
4.4.1 Learning by Analogy or Learning by Rules 89
4.4.2 Partial Regularity and History 92
4.5 Some Purported Evidence for the WR Model 94
4.5.1 Error Rate 94

4.5.2 The role of input frequency . . . 96
4.5.3 The postulation of the -d rule . . . 96
4.5.4 Gradual improvement . . . 99
4.5.5 Children's judgment . . . 99
4.5.6 Anecdotal evidence . . . 100
4.5.7 Adult overregularization . . . 100
4.5.8 Indecisive verbs . . . 101
4.5.9 Irregulars over time . . . 101
4.5.10 Corpus statistics . . . 102
4.6 Conclusion . . . 102

5 Internal and External Forces in Language Change 107
5.1 Grammar Competition and Language Change . . . 109
5.1.1 The role of linguistic evidence . . . 109
5.1.2 A Variational Model of Language Change . . . 111
5.2 The Loss of V2 in French . . . 115
5.3 The Erosion of V2 in Middle English . . . 118
5.3.1 Word order in Old English . . . 119
5.3.2 The southern dialect . . . 120
5.3.3 The northern dialect and language contact . . . 122
5.4 Conclusion . . . 123

6 A Synthesis of Linguistic Knowledge and Learning 125

7 Bibliography 128

Chapter 1

The Study of Language and Language Acquisition

> We may regard language as a natural phenomenon — an aspect of his biological nature, to be studied in the same manner as, for instance, his anatomy.
>
> Eric H. Lenneberg
> *Biological Foundations of Language* (1967, vii)

1.1 The Naturalistic Approach to Language

Fundamental to modern linguistic inquiry is the view that human language is a natural object: our species-specific ability to acquire a language, our tacit knowledge of the enormous complexity in language, and our capacity to use language in free, appropriate, and infinite ways, are all attributed to some properties of the natural world, our brain. This position need not be defended, if one considers the study of language is of any scientific interest whatsoever.

It follows then, as in the study of biological sciences, linguistics aims to identify the abstract properties of the biological object under study, namely, human language, and the mechanisms that underly its organization. This is a goal set in the earliest statements on modern linguistics, Noam Chomsky's *The Logical Structure of Linguistic Theory* (1955).

Consider the famed pair of sentences:

(1) a. Colorless green ideas sleep furiously.

b. * Furiously sleep ideas green colorless.

Neither sentence has even a remote chance to be encountered in natural discourse, yet every speaker of English can perceive their differences: while they are both thoroughly meaningless, (1a) is grammatically well-formed, whereas (1b) is not. To give "a rational account of this behavior, i.e., a theory of the speaker's linguistic intuition ... is the goal of linguistic theory" (Chomsky 1955/1975: p95) – in other words, a psychology, and ultimately, biology, of human language.

Once the position, dubbed *biolinguistics* (Jenkins 1999, Chomsky 1999), is accepted, it immediately follows that language, just as all other biological objects, ought to be studied following the standard methodology in natural sciences (Chomsky 1980, 1986, 1995a). The postulation of innate linguistic knowledge, the Universal Grammar (UG), is a case in point.

One of the major motivations for innateness of linguistic knowledge comes from the Argument from the Poverty of Stimulus (APS) (Chomsky, 1980: p35). A well-known example concerns the *structure dependency* in language syntax and children's knowledge of it in the absence of learning experience (Chomsky 1975, Crain and Nakayama 1987). Forming a question in English involves inversion of the auxiliary verb and the subject:

(2) a. Is Alex *e* singing a song?

b. Has Robin *e* finished reading?

It is important to realize that exposure to such sentences underdetermines the correct operation for question formation. There are many possible hypotheses compatible with the patterns in (2):

(3) a. front the first auxiliary verb in the sentence

b. front the auxiliary verb that is most closely follows a noun

c. front the second word in the sentence

d. ...

The correct operation for question formation is, of course, structure dependent: it involves parsing the sentence into structurally organized phrases, and fronting the auxiliary that follows *the first noun phrase*, which could be arbitrarily long:

(4) a. Is [the woman who is sing] *e* happy?

b. Has [the man that is reading a book] *e* had supper?

Hypothesis (3a), which can be argued to involve simpler mental computation than the correct generalization, unfortunately yields erroneous predictions:

(5) a. * Is [the woman who *e* singing] is happy?

b. * Has [the man that *e* finished reading] has finished supper?

But children don't go astray like the creative inductive learner in (3). They stick to the correct operations from as early as we one test them, as Crain and Nakayama (1987) did using elicitation tasks. The children were instructed to "Ask Jabba if the boy who is watching Mickey Mouse is happy", and no error of the form in (5) was found.

Though sentences like those in (4) may serve to disconfirm hypothesis (3a), they are very rarely if ever encountered by children in normal discourse,[1] not to mention the fact that each of the other incorrect hypotheses in (3) will need to be ruled out by disconfirming evidence. Here lies the logic of the APS:[2] if we know X, and X is underdetermined by learning experience, then X must be innate. The conclusion is then Chomsky's (1975: p33): "the child's mind ... contains the instruction: Construct a structure-dependent rule, ignoring all structure-independent rules. The principle of structure-dependence is not learned, but forms part of the conditions for language learning."

The naturalistic approach can also be seen in the evolution of linguistic theories via successive refinement and revision of ideas as their conceptual and empirical flaws are revealed. For example, the 1960's language-particular and construction-specific transformational rules, while descriptively powerful, are inadequate when viewed in a biological context. The complexity and unrestrictiveness of rules made the acquisition of language wildly difficult: the

[1] In section 3.2, we will use corpus statistics from Legate (1999) to make this remark precise, and to address some recent challenges to the APS raised by Sampson (1989) and Pullum (1996).

[2] See Crain (1991) for several similar cases, and numerous others in the child language literature.

learner would have a vast (and perhaps an infinite) space of hypotheses to entertain. The search for a plausible theory of language acquisition, coupled with the discoveries in comparative studies, led to the Principles and Parameters (P&P) framework (Chomsky 1981), which suggests that all languages obey a universal (and hence putatively innate) set of tightly constrained principles, whereas variations across constructions and languages — the space of language acquisition — are attributed to a small number of parametric choices.

The present dissertation is a study of language development in children. From a biological perspective, the development of language, as the development of any other organic systems, is an interaction between internal and external factors; specifically, between the child's internal knowledge of linguistic structures and the external linguistic experience he receives. Drawing insights from the study of biological evolution, I will put forth a quantitative model of language acquisition that make this interaction precise, by embedding a theory of knowledge, the Universal Grammar, into a theory of learning from experience. In particular, I will advance the idea that language acquisition should be modeled as a population of "grammars", competing to match the external linguistic experiences, much like in a natural selection process. The justification of this outrageous idea will take the naturalistic approach just as in the justification of innate linguistic knowledge: I will provide evidence — conceptual, mathematical, and empirical, and from a number of independent areas of linguistic research, including the acquisition of syntax and morphophology, and historical language change – to convince you that it has to be the case.

But before we dive into details, some methodological remarks on the study of language acquisition.

1.2 The Structure of Language Acquisition

At a most abstract level, language acquisition can be represented as follows:

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

A learning function or algorithm $\mathcal{L}$ maps the initial state of the learner, S_0, to the terminal state S_T, on the basis of experience E in the environment. Language acquisition research attempts to give an explicit account of this process.

1.2.1 Formal Sufficiency

The acquisition model must be *causal* and *concrete*. Explanation of language acquisition is not complete with a mere description of child language, no matter how accurate or insightful, without an explicit account of the mechanism responsible for how language develops over time, the learning function $\mathcal{L}$. It is often claimed in the literature that children just "pick up" their language, or some parameter is somehow set correctly, or children's linguistic competence is identical to adults. Statements as such, if devoid of a serious attempt at some learning-theoretic account of *how* the child manages, profess more irresponsibility than ignorance.

The model must also be *correct*. Given reasonable assumptions about the linguistic data, the duration of learning, the learner's cognitive and computational capacities, and so on, the model must be able to attain the *terminal* (rather than target) state of linguistic knowledge S_T comparable to that of a normal human learner. The correctness of the model must be verified by mathematical proof, computer simulation, or any other kind of rigorous demonstration. This requirement has traditionally been referred as the *learnability condition*, which unfortunately carries some misleading connotations. For example, the influential Gold paradigm of identification in the limit (1967) requires the learner to converge on to the target grammar in the environment. However, this position has little empirical content.[3]

First, language acquisition is the process in which the learner forms an *internalized* knowledge (in his mind), an I-language (Chomsky 1986): there is no external target of learning, hence, no "learnability" in the traditional sense. Section 1.2.2 below documents evidence that child language and adult language appear to be sufficiently different such that language acquisition can not be viewed as recapitulation or approximation of an external learning target. Second, in order for language to change, the terminal state attained by children must be different from their ancestors – which requires the learnability condition *fail*. In particular, as we shall see in Chapter 5 empirical cases where learners do not converge to any unique "grammar" in the informal sense of "English" or "German", but rather a combination of multiple grammars. Language change is a result of changes in this

[3]I am indebted to Noam Chomsky for discussion of this issue over the years.

kind of grammar combinations. A correct model must be able to accommodate all these varied outcomes of language acquisition.

1.2.2 Developmental Compatibility

A model of language acquisition is, after all, a model of reality: it must be compatible with what has been discovered about children's language.

Essential to this requirement is the *quantitativeness* of the model. This condition echoes the quantitative approach that has become dominant in theoretical language acquisition over the past two decades — it is no coincidence that the maturation of theoretical linguistics and the availability of large scale child language databases (MacWhinney & Snow 1986) took shape in the same time.

When studying the problem of learning, it is often revealing to make quantitative comparisons of what is learned and how the learner learns it: in our case, quantitative measures of child language and those of adult language. Here many intriguing disparities surface. A few examples illustrate this observation and the challenge it poses.

It is now known that some aspects of the grammar are acquired successfully at a remarkably early age. The placement of finite verbs in French matrix clauses is such an example.

(7) Jean voit *souvent/pas* Marie.
Jean sees *often/not* Marie.
'John *often* sees/does *not* see Marie.'

French, in contrast to English, places finite verbs in a position preceding sentential adverbs and negations. Although sentences like (7), indicative of this property of French, are quite rare in adult-to-child speech (7-8%, in our estimate based on CHILDES), French children, since as early as can be tested (the 18th month; Pierce 1989), almost never deviate from the correct form.[4]

In contrast, some very robustly attested patterns in adult language emerge rather late in children. The best known example is perhaps the phenomenon of subject drop. Children

[4]This discovery has been duplicated in a number of languages that have similar properties; see Wexler (1994) for a survey.

learning English, German, and other languages that require the presence of a grammatical subject often produce sentences as in (8):

(8) a. (I) help Daddy.

b. (He) dropped the candy.

Subject drop appears in up to 30% of all sentences around 2;0, and it is not until around 3;0 they start using subjects at adult level (Vilian 1991, Wang et al. 1992), in striking contrast to adult language, which uses subject in almost all sentences.

Even more interestingly, children often produce utterances that are virtually *absent* in adult speech. In what is descriptively known as the Optional Infinitive (OI) stage (Pierce 1989, Weverink 1989, Wexler 1994), children acquiring some languages produce a significant number of sentences where matrix verbs are non-finite. (9) is an OI example from child Dutch (Weverink 1989):

(9) pappa schoenen wassen
daddy shoes to-wash
'Daddy washes shoes'.

Non-finite root sentences like (9) are ungrammatical in adult Dutch and thus appear very infrequently in the linguistic evidence. Yet OI sentences are robustly used by children for an extended period of time, before they gradually disappear by 2;6 or later.

These quantitative disparities between child and adult language pose considerable difficulty for inductive data-driven learning approaches to language acquisition. The problem is, as pointed out by Fodor & Pylyshyn (1988), a statistical language learning model with prior knowledge (such as the UG) can do no more than recapitulating the statistical distribution of the input data. It is therefore unclear how a statistical learning model can duplicate the developmental patterns in child language: early acquisition despite sparse frequency (French verb placement), late acquisition despite overwhelming frequency (English subject drop), and robust use despite virtual absence (OI).[5]

[5]Nor is there any obvious extralinguistic reason why the early acquisitions are intrinsically "simpler" to learn than the late acquisitions. For instance, both the obligatory use of subject in English and the placement of finite verbs before/after negation and adverbs involve a binary choice.

Even with the assumption of innate UG, which can be viewed a kind of prior knowledge, it is not clear how such quantitative disparities can be explained. As will be discussed in Chapter 2, previous formal models of acquisition in the UG tradition in general have not even begun to address these questions. The model developed in this study intends to fill this gap.

Finally, quantitative modeling is important to the development of linguistics at large. At the foundation of every "hard" science is a formal model in which quantitative data can be explained and quantitative predictions can be made and checked. Biology did not come to age until its two pillars of biological sciences, Mendelian genetics and Darwinian evolution, were successfully integrated in the mathematical theory of population genetics, where evolutionary change can be explicitly and quantitatively expressed by its internal genetic basis and external environmental conditions.[6] It is certainly desirable if the interplay between the internal linguistic knowledge and the external linguistic experience can be quantitatively modeled under some appropriate framework.

1.2.3 Explanatory Continuity

As it is obvious that child language differs from adult language, it is absolutely necessary for an acquisition to make some choices on explaining such differences. The condition of *Explanatory Continuity* proposed here imposes some restrictions, or, to be more precise, heuristics on making these choices.

Explanatory Continuity is an instantiation of the well-known Continuity Hypothesis (MacNamara 1982, Pinker 1984). The Continuity Hypothesis says, without evidence to the contrary, children's cognitive system is assumed to be identical to adults'. If child and adult languages differ, two possibilities may be at issue:

(10) a. Children and adults differ in linguistic performance.

b. Children and adults differ in grammatical competence.

An influential view holds that child competence is *identical* to adult competence (Pinker 1984), which necessarily leads to a performance-based explanation for child acquisition.

[6]See Lewontin (1996) and Maynard Smith (1989) for two particularly insightful expositions on population genetic theories.

There is no question that (10a) is, at some level, true: children are more prone to performance errors than adults, as their memory, processing, and articulation capacities are still in development. But there are reasons to prefer competence-based explanations. Parsimony is an obvious one. By definition, performance involves the interaction between the competence system and other cognitive/perceptual systems. In addition, competence is about the only system in linguistic performance for which theoretical understanding is at some depth. This is partially because grammatical competence is to a large degree isolated from other cognitive systems – the autonomy of syntax – and is thus more directly accessible for investigation. The tests used for competence studies, often in the form of native speaker's grammatical intuition, can be carefully controlled and evaluated. Finally, and empirically, child language differs from adult language in very specific ways, which do not seem to follow from any general kind of deficit in children's performance. For example, it has been shown that there is much data in child subject drop that do not follow from performance-based explanations; see Hyams and Wexler (1993), Roeper and Rohrbarher (1994), Bromberger and Wexler (1995), etc. In Chapter 3, we will present additional developmental data from several studies of children's syntax to show the insufficiency of performance-based approaches.

If we tentatively reject (10a) as (at least) a less favorable general research strategy, we must rely on (10b) to explain child language. But exactly *how* is child competence different from adult competence? Here again are two possibilities:

(11) a. Child competence and adult competence are qualitatively different.

b. Child competence and adult competence are quantitatively different.

(11a) says that child language is subject to different rules and constraints from adult language. For example, it could be that some linguistic principle operates differently in children from adults, or a piece of grammatical knowledge is absent in younger children but become available as a matter of biological maturation (Gleitman 1981, Felix 1987, Borer and Wexler 1987).

I wish to stress that there is nothing unprincipled in postulating a discontinuous competence system to explain child language. If children systematically produce linguistic expressions that defy UG (as understood via adult competence analysis), we can but only conclude that their language is governed by different laws. in the absence of a concrete

theory of how linguistic competence matures, (11a) runs the risk of "anything goes". It must therefore remain a last resort only when it has been proven hopeless to follow the direction of (11a) by explaining child language with adult competence, for which we *do* have concrete theories.[7] More specifically, we must not confuse the difference between child language and adult language with the difference between child language and Universal Grammar. That is, while (part of) child language may not fall under the grammatical system he eventually attains, it is possible that it falls under some *other* grammatical system allowed by UG. (Indeed, this is the approach advocated in the present study.)

This leaves us with (11b), which, in combination with (10b), gives a strongest realization of the Continuity Hypothesis: that child language is subject to same principles and constraints in adult language, and that every utterance in child language is *potentially* an utterance in adult language. The difference between child and adult languages is due to differences in the *organization* of a continuous grammatical system. This position further splits into two directions:

(12) a. Child language reflects a *unique* potential adult language.

b. Child grammar consists of a *collection* of potential adult languages.

(12a), the dominant view ("triggering") in theoretical language acquisition will be rejected in Chapter 2. Our proposal takes the position of (12b): child language in development reflects a combination of possible grammars allowed by UG, only some of which are eventually retained when language acquisition ends. It will be elaborated in the rest of this dissertation, where we examine how it measures up against the criteria of formal sufficiency, developmental compatibility, and explanatory continuity.

1.3 A Roadmap

This dissertation is organized as follows.

[7]This must be determined case by case, although it is often the case that when maturational accounts have been proposed, non-maturational explanations of the empirical data have not been conclusively ruled out, and hence superior on principled ground. For example, Borer and Wexler's proposal (1987) that certain A-chains mature have been called into question by many researchers (Pinker, Lebeaux, and Frost 1987, Demuth 1989, Crain 1991, Fox, Grodzinsky, and Crain 1995, etc.).

Chapter 2 first gives a critical review of the formal approaches to language acquisition. After an encounter with the populational and variational thinking in biological evolution, which indeed inspired this dissertation, we propose to model language acquisition as a population of competing grammars, whose distribution changes in response to the linguistic evidence presented to the learner. We will give a precise formulation of this idea, and study its formal/computational properties with respect to the condition of *formal sufficiency.*

Chapters 3 subjects the model to the test of *developmental compatibility* by looking at the acquisition of syntax. First, cross-linguistic evidence will be presented to showcase the model's ability to make quantitative predictions based adult-to-child corpus statistics. In addition, a number of major empirical cases in child language, including the acquisition of word order in a number of languages and the subject drop phenomenon will be reexamined under the present framework to demonstrate the reality of co-existing and competing grammars in child language.

Chapter 4 applies the model to yet another major problem in language acquisition, the learning of English irregular verbs. It will be shown that irregular verbs are organized into classes, each of which is defined by a special phonological rule, and that learning an irregular verb involves the competition between the designated special rule and the default -ed rule. Again, quantitative predictions are made and checked against children's performance on irregular verbs. Along the way we will develop a critique of Pinker and his colleagues' *Words and Rule* model, which holds that irregular verbs are individual and directly memorized as associated pairs of root and past tense forms.

Chapter 5 extends the acquisition model to the study of language change. The quantitativeness of the acquisition model allows one to view language change as the change in the distribution of grammars in successive generations of learners, which can directly related to the statistical properties of historical texts in an evolving dynamical systems. We apply the model of language change to explain the loss of Verb Second in Old French and Old English.

Chapter 6 concludes with a discussion on the implications of the acquisition model in a broad context of linguistic research.

Chapter 2

A Variational Model of Language Acquisition

One hundred years without Darwin are enough.

H. J. Muller (1959) at the centennial
celebration of *On the Origin of Species*.

It is a simple observation that young children's language is different from adults'. However, this simple observation raises quite 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 (2), provides a useful framework to characterize approaches to language acquisition:

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

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 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 entertain in principle. Almost all theories of acquisition in the UG-based approach can be referred to as *transformational learning* models, borrowing a term from evolutionary biology (Lewontin 1983), in the sense that 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. In the remainder of this dissertation, I 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 domain-neutral learning theories from traditional psychology.

2.1 Against Transformational Learning

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

(14) a. Formal sufficiency.

b. Developmental compatibility.

c. Explanatory continuity.

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

In recent years, the GSL approach to language acquisition has (re)gained popularity in cognitive sciences and computational linguistics; see, for example, Bates and Elman (1996), Seidenberg (1997), among many others. The GSL approach claims to assume little about

the learner's initial knowledge of language.[1] 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 and Elman 1996) from the demonstration that infants are capable of using statistical information to identify word boundaries (Saffran, Aslin, and Newport 1996; *inter alia*).

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 very rudimentary form of the formal sufficiency condition — although there is still much debate on whether this project has been at all successful (e.g., Marcus 1998, *inter alia*). Much more effort has gone into the learning of irregular verbs, starting with Rumelhart and McClelland (1986) and followed by numerous others,[2] which prompted the reviewer of the connectionist manifesto *Rethinking Innateness* (Elman et al. 1996) to remark that connectionist modelings makes one feel as if developmental psycholinguistics is only about "development of the lexicon and past tense verb morphology"(Rispoli 1999: p220), aside from the fact that, as we shall see in Chapter 4, much of the complexity in past tense verb morphology still remains uncovered and unexplained.

Nothing can be said unless the GSL approach faces the challenges from the cross-linguistic study of child language. There is reason to believe that this challenge is formidable. As remarked in section 1.2.2, child language and adult language display significant disparities in statistical distributions, which makes a naive GSL approach (learning by approximation) incoherent. What the GSL approach has to do, then, is to find a empiricist (learning-theoretic) alternative to the learning "biases" introduced by innate UG.[3]

We thus focus our attention on the other leading approach to language acquisition, one which is most closely associated with generative linguistics. We will not review the argument for innate linguistic knowledge; see section 1.1 for a simplest but most convincing

[1]However, a recent study by Marcus (1998) shows that in fact many hidden assumptions about the object of learning are built in connectionist models, a most visible advocate of the GSL approach.

[2]Pinker (1999: p302) lists twenty-five major connectionist studies on irregular verbs.

[3]The manner in which UG provides learning biases (priors) will be made precisely by the present model.

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 (Chomsky 1965, Wexler and Culicover 1980, Berwick 1985, Hyams 1986, Dresher and Kaye 1990, Gibson and Wexler 1994, etc.) is what 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 and Culicover 1980, Berwick 1985) that when presented with a sentence that the learner is unable to analyze, an appropriate transformational rule is added to the current hypothesis (a set of transformational rules). 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 (Gibson and Wexler 1994). In Gibson and Wexler's Triggering Learning 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-none 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, 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 the Gibson and Wexler triggering model has a number of formal problems; see Berwick and Niyogi (1996), Frank and 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.[4] By analyzing the triggering model as a Markovian process in a finite space of grammars, Berwick and Niyogi (1996) have demonstrated the pervasiveness of local maxima in Gibson and Wexler's (very small) 3-parameter space. Gibson and 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), via exhaustive search in a computer implementation of the triggering model, shows that in a linguistically realistic 12-parameter space, 2,336 of the 4,096 grammars are still not learnable even 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.[5]

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 grammars. For example, a sentence with an overt thematic subject is ambiguous between an English type grammar, which obligatorily uses subjects and an Italian or Chinese type grammar, which optionally uses subjects. When ambiguous evidence is presented, it may select any of the grammars compatible with the evidence and subsequently be led to local maxima and unlearnability. To resolve the ambiguity problem, Fodor (1998) suggests 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 and Wexler's 3-parameter space, it is doubtful that such an optimistic expectation holds for all natural language grammars (Clark and 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.

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

[5]Niyogi and 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. However, empirical facts from diachronic studies suggest otherwise; see Chapter 5 for discussion.

Lastly, the robustness of the triggering model has been called into question. As pointed out by Osherson, Weinstein, and Stob (1982) and Valian (1990), even a small amount of noise can mislead the triggering-like transformational models to converge on a wrong grammar. In a most extreme form, if the *last* sentence the learner hears just before language acquisition stops happens to be noise, the learning experience during the entire duration of language acquisition goes wasted in vain. This scenario is by no means an exaggeration when a realistic learning environment is taken into account. Actual linguistics environments are hardly uniform with respect to a single idealized grammar. For example, Weinreich, Labov, and Herzog (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, for example, *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 (Randall 1990, Valian 1990).

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 and Berwick 1996)), the difficulty posed by the developmental compatibility condition seems far more serious. In the triggering model, and in fact in *all* TL models, the learner at any time is be identified with a single grammar. If such models are at all relevant to the explanation of child language, the following predictions are inevitable:

(15) 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 its linguistic expressions should be observed.

To the best of my knowledge, there is, in general, no developmental evidence for either (15a) or (15b).

A good test case is again the null subject (NS) phenomenon, on which there is a large body of cross-linguistic literature. First, let's examine the prediction in (15a), the consistency of child language with respect to a single grammar defined in the UG space. Working in the P&P framework, Hyams (1986) suggests that English child NS results from mis-setting 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 (adult-level performance), the subject drop rate is only 31% for American children in the same age group. The 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, Lillo-Martin, Best, and Levitt (1992) show that while subject drop rate is only 26% for American children during the NS stage (2;0-3;0),[6] Chinese children in the same age group drop subjects in 55% of all sentences (also adult-level performance). 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 3.3.2 for additional discussion). This is again incorrect: Wang et al. (1992) find that Chinese children drop objects in 20% of sentences containing objects, and English children, only 8%. These comparative studies conclusively demonstrate that subject drop in child English cannot be identified with any single adult grammar.

Turing now to the triggering models' second prediction for language development (15b), we expect to observe abrupt changes in child language as the learner switches from one grammar to another. However, Bloom (1993) found no sharp change 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. This is contrary to the parameter re-setting proposal of Hyams

[6]This figure is considerably lower than those reported elsewhere in the literature, e.g., Bloom (1993), Hyams and Wexler (1993). However, there is good reason to believe 26% 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 have been acceptable in adult English; see Phillips (1995) for an extended discussion on the counting procedure.

and Wexler (1993), and the triggering model in general. In section (3.1), we will show that for Dutch children, 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 under 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 1993, Elman et al. 1996), forcing many researchers to the position 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:

(16) a. Children and adults differ in linguistic performance.

b. Children and adults differ in grammatical competence.

The performance deficit approach (16a) is often couched 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 (non-linguistic) perceptual and cognitive capacities that are still underdeveloped at a young age (Pinker 1984, Bloom 1990, 1993 Gerken 1991, Valian 1991, *inter alia*).

The competence deficit approach (16b) 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.[7] 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

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

Tense/Agr node in children's syntactic representation (Wexler 1994): the Tense/Agreement features are missing in young children during the RI stage. Another influential proposal in Rizzi's (1994) *Truncation Hypothesis*, which 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 survey of some recent proposals along these lines of theorizing.

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 approaches. More empirically, as we shall see in Chapter 3, the imperfection perspective on child language leaves many developmental patterns unexplained. To give a quick preview, we will show that when English children drop subjects in wh questions, they only do so in adjunct (*where*, *how*) questions, but not in argument (*who*, *what*) questions: a categorical asymmetry not predicted by either performance or competence based approach. We will show the robust use (approximately 50%) of V1 patterns in V2 acquisition: identifying child competence with adult competence under the Continuity Hypothesis, or claiming early setting of the V2 parameter (Poeppel and Wexler 1993, Wexler 1998), is indistinguishable from saying that children use the V2 grammar at chance.

To end this very brief review of leading approaches to language acquisition, notice that there is something too sensible to dismiss in both GSL models and UG-based transformational learning. On the one hand, the gradualness of language development seems most naturally captured in the statistical terms of the GSL models. And on the other, the restrictiveness of natural languages revealed by child language research and comparative linguistics must play a crucial role in constraining the child's hypothesis space and the learning process. In the rest of this chapter, I propose a new approach that combines the virtues of both models: 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. I believe that 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 in Darwin's days. As pointed out by Ernst Mayr on many occasions, particularly in *Animal Species and Evolution* (1963), 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 Richard 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: 65-66)

The message embedded in the Darwinian variational thinking is a profound one for making scientific observations. 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 to be 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 grammars, the principled variations in human language.

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

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 subscribe to the assumption of the P&P framework, i.e., there is only a finite number of possible human grammars. We also adopt the strongest version of continuity hypothesis, the default, which says, unless proven wrong, the 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 confidence the learner associates with G_i. 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.[8] In particular, in an idealized environment where *all* linguistic expressions are generated by a target grammar T, which belongs to the finite UG space, 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:

(17) 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

Metaphorically speaking, the learning hypotheses – the grammars defined by UG — *compete*: grammar that succeed to analyze a sentence are rewarded and grammars 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 grammar G_1, the target grammar used in the environment, and G_2, the competitor, with associated weights of p_1 and p_2 respectively. Suppose that 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 the learner produces will look like:

(18) 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.

[8]This does not mean that learning necessarily converges to a single grammar; see (23) below.

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

(19) Immediate in acquisition:

$S_{G1}, S_{G1}, S_{G2}, S_{G1}, S_{G1}, S_{G1}...$

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:

(20) Completion of acquisition:

$S_{G1}, S_{G1}, S_{G1}, S_{G1}, S_{G1}, S_{G1}, ...$

Of course, grammars do not *actually* compete with each other: the competition metaphor only serves to illustrate (a) the grammars' co-existence, and (b) their differential representation in the learner's language faculty. Neither does the learner plays God by supervising the competition of the grammars and selecting the winners.[9] I must 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 two-fold. On the one hand, successful language acquisition is possible, barring pathological cases, irrespective of "general intelligence"; on the other, we just 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 that happens to be selected with its associated weight/probability. He does not make active changes to the *grammar* (as in the triggering model), or reorganize his grammar space but simply updates the weight of the single grammar selected for analysis and moves on.

[9]This is contrasted with a similar model of acquisition (Clark 1992), in which the learner is viewed as a genetic algorithm which explicitly evaluates grammar fitness.

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, in a special case, can be interpreted as parsability (Wexler and Culicover 1980, Berwick 1985). It is worth noting that this rather narrow definition of analyzability does not affect the formal properties of the model: *any* notion of analyzability, as long as it's well-defined, is compatible with the learning model.[10] The choice of parsability obviously eases the evaluation of grammars using a linguistic corpus.

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 changes in the weights of grammars in response to the sentences successively presented to the learner. There are many possible instantiations of competition-based learning.[11] Consider the one in (21):

(21) Given an input sentence s, the child

a. with the probability p_i, selects a grammar G_i

- if $G_i \rightarrow s$ then $\begin{cases} p'_i = p_i + \gamma(1-p_i) \\ p'_j = (1-\gamma)p_j \qquad \text{if } j \neq i \end{cases}$
- if $G_i \not\rightarrow s$ then $\begin{cases} p'_i = (1-\gamma)p_i \\ p'_j = \frac{\gamma}{N-1} + (1-\gamma)p_j \quad \text{if } j \neq i \end{cases}$

(21) is the Linear reward-penalty (L_{R-P}) scheme (Bush and Mosteller 1951, 1958), one of the earliest 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, Bower, and Crothers (1965) for a review.

Does the employment of a general-purpose learning model in the behaviorist tradition, the L_{R-P}), signal a return to the dark ages? Absolutely not. In competition learning models,

[10] Perhaps along the lines suggested by Tom Roeper in many places that grammar/data match ought to be multi-dimensional, including phonological, syntactic, as well as pragmatic factors.

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

what is crucial is the constitution of the hypothesis space. In the original L_{R-P} scheme, the hypothesis space consists of simple responses 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 incoherent that human language acquisition can be equated to data driven learning without prior knowledge. And, as will be shown in Chapters 3 and other places, in order to account for child language development, one will need make reference to specific characterization of UG supplied by linguistic theories. What we advocate here is simply a plea to pay attention to the actual mechanism of language acquisition, 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 (21).

2.3.1 Asymptotic Behaviors

In any competition process, some measure of fitness is required. Adapting the formulation of Bush and Mosteller (1958), we may define:

(22) 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 that the grammar G_i is incompatible with. Notice that penalty probability is an *intrinsic* property of a grammar relative to a fixed linguistic environment E, from which input sentences are drawn.

For example, consider a Germanic V2 environment, in which all sentences have the main verb in the second constituent position. A V2 grammar, of course, has the penalty probability of 0.[12] An English type SVO grammar, although not compatible with all V2 sentences,

[12]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, with desirable consequences in

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 definition, fixed.

It is crucial to realize that penalty probability is an extensionally defined property of grammars, used only in the formal analysis of the learning model. It is not a component of the learning process. For example, the learner need 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 (21) makes clear.

The asymptotic properties of models like (21) have been extensively and rigorously studied in both mathematical psychology (Norman 1972) and machine learning (Narendra and Thathachar 1989, Barton and Sutton 1998). For simplicity but without loss of generality, suppose that there are two grammars in the population, G_1 and G_2, and they are associated with penalty probabilities of c_1 and c_2 respectively. If the learning rate γ is sufficiently small, that is, the learner does not alter its "confidence" in grammars too radically, one can show (see Narendra and Thathachar 1989:162-165) that the asymptotic distributions of $p_1(t)$ and $p_2(t)$ will be essentially normal and can be approximated as follows:

(23) **Theorem**:

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

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

(23) 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.

---

the reduction of computational cost.

2.3.2 Stable Multiple Grammars

Recall from section 2.1.1 that realistic linguistic environments are usually *non-homogeneous*: linguistic expressions 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 (23) 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), and may offer a way to integrate generative linguists' idealized grammars with the study of language variation and use in sociolinguistic research (see, for example, the collection in Sankoff (1978)).

A most radical case of non-homogeneous environment is one in which speakers of two different "languages" come into contact. It follows that the learner forms a stable combination of two grammars. This conclusion is confirmed in the diachronic studies by Kroch and his colleagues (Kroch 1989, Pintzuk 1991, Santorini 1992, Kroch and Taylor 1997). They show that during the course of language change, speakers are best viewed as accessing stable multiple grammars, rather than as a single grammar supplemented with additional assumptions.

One generation of "multilingual" speakers produces non-homogeneous expressions, which constitutes yet another non-homogeneous environment for the generation to follow. This iterative process can be studied as an evolving dynamical system. In Chapter 5, we extend the present model of language acquisition to a model of language change. We show that a combination of grammars as the result of acquisition, while stable in a synchronic generation of learners, may not be diachronically stable. We will derive precise 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 others grammars in the population have positive penalty probabilities. It is easy to see from (23) that the p_1 converges to 1, with the competing grammars eliminated. Thus, the variational model meets the traditional learnability condition.

Empirically, perhaps the most important feature 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:

$$
\begin{aligned}
E[\Delta p_1] &= p_1\gamma(1-p_1)+ && \text{with Pr. } p_1,\ G_1 \text{ is chosen and } G_1 \rightarrow s \\
&\quad\ p_2(1-c_2)(-\gamma)p_1+ && \text{with Pr. } p_2(1-c_2),\ G_2 \text{ is chosen and } G_2 \rightarrow s \\
&\quad\ p_2c_2\gamma(1-p_1) && \text{with Pr. } p_2c_2,\ G_2 \text{ is chosen but } G_2 \not\rightarrow s \\
&= c_2\gamma(1-p_1)
\end{aligned}
\tag{24}
$$

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 about 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, (24) allows us to make longitudinal predictions about language development that can be verified against actual findings. In Chapter 3, we do just that.

A disclaimer, or rather, a confession, is in order. We am in fact 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: co-existing hypotheses in gradual competition, as schematically illustrated in (17). 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 (23), which is needed to accommodate the empirical fact of linguistic

variation that is particularly clear in language contact and change. That is, stable co-existing grammars can be retained in mature speakers when the linguistic evidence cannot be exclusively attributed to a single grammar.[13] Again, this property of the model will prove crucial when we consider the independent facts from language change in Chapter 5.

2.3.3 Unambiguous evidence

The theorem in (23) states that in the variational model, convergence to the target grammar is guaranteed if all competitor grammars have positive penalty probabilities. Obviously, one way to ensure this is to assume the existence of unambiguous evidence (Fodor, 1998): sentences that are compatible with only the target grammar but not with any other grammar. While the general existence of unambiguous evidence has been questioned (Clark and Roberts 1993), we shall point out that 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 grammars:

(25) a. Dutch: SVO, XVSO, OVS

b. Arabic: SVO, XVSO

c. English: SVO, XSVO

d. Irish: VSO, XVSO

e. Hixkaryana: OVS, XOVS

The grammars in (25) are followed by some of the matrix sentences word orders they can generate/analyze.[14] Observe that none of the patterns in (25a) *alone* could distinguish Dutch from the other four human grammars, as each of them is compatible with certain

[13]The model in Yang and Gutmann (1999), in contrast, converges on to the *fittest* grammar, even if it is not 100% compatible with the input data.

[14]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. It is perhaps not true that V2 does not have direct unambiguous evidence, if we relax the degree-0 constraint on acquisition, since children can use embedded clause as evidence for underlying word order (Roeper 1973). The use of the V2 example here is to illustrate the acquisition of a grammar in the absence of unambiguous cues. In addition, what Lightfoot (1999) concludes to be unambiguous cue for the V2 grammar, the OVS pattern, is also what we regard to be the *effective* unambiguous evidence; see section 3.1.2 for detail.

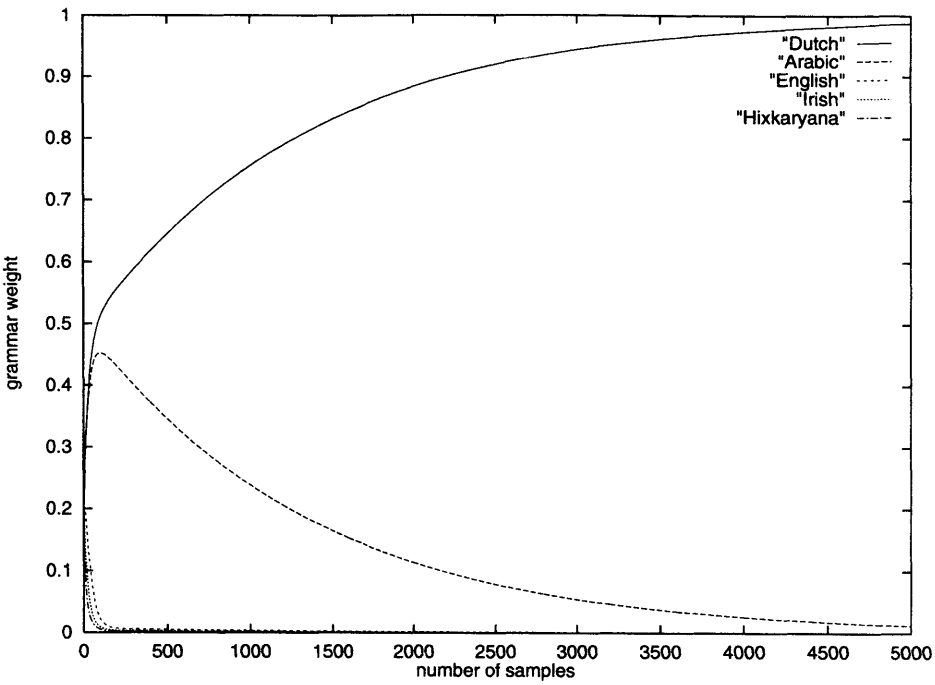

Figure 2-1: The convergence to the V2 grammar in the absence of unambiguous evidence.

V2 sentences. Specifically, based on the input evidence a Dutch child (Hein) received, we found that 64.7% are SVO patterns, followed by XVSO patterns at 34% and only 1.3% OVS patterns.[15] Most notably, the Arabic type grammar, which allows SVO and VSO alternations (Greenberg 1963: Universal 6; Fassi-Fehri 1993), is compatible with 98.7% of V2 sentences.

Despite of 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 (25) will be punished. The V2 grammar, however, is never punished. The theorem in (23) thus ensures the learner's convergence to the target V2 grammar. The competition of grammars is illustrated below, based on a computer simulation.[16]

[15] I am grateful to Edith Kaan for her assistance in the counting.

[16] I should add that the time X-axis, the number of example sentences the learner encounters, only serves to show the *relative* changes in grammar weights. Exactly how much time is required for the V2 grammar to win out will depend on many unknown parameters such as the learning rate γ.

2.4 Learning Grammars in a Parametric Space

It should be clear that the variational model developed in the preceding sections is entirely theory neutral. The model only requires a finite and non-arbitrary space of hypotheses – a conclusion which is accepted by many of today's linguists — no other particular properties of the UG need to be assumed.[17] We now situate the learning model in a realistic grammar space, the P&P framework.

2.4.1 Parameters Make Learning Efficient

So far we have been treating competing grammars as individual entities: we have not taken into account the structure of the grammar space. Despite that the fact that the two grammar result in (23) 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. If, according to some estimates (Clark 1992), 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 cost of learning. Suppose there are n parameters, $\alpha_1, \alpha_2, ..., \alpha_n$, and each grammar can be represented as a vector of 0's and 1's. We will be able to represent the 2^n grammar weights as a n-dimensional vector of real numbers between [0, 1]: (w_1, w_2, ..., w_n), where w_n denotes the weight of the parameter α_i setting to the value [+] (or [−]). The values of m parameters will be independently selected according to the parameter weights. Then learner can "compile" a parameter vector (a vector of 0's and 1's) into a usable grammar (Fong 1991, Clark 1992):

(26) a. Select parameter values independently

b. Compile a parameter vector to analyze the incoming sentence

c. Update the parameter values accordingly

[17] Different theories of UG will of course yield different generalizations: when situated into a theory-neutral learning model, they will presumably make different developmental predictions. The present model can then be used as an independent procedure in evaluating linguistic theories. See Chapter 6 for additional discussion.

Now a problem of *parameter interference* immediately arises.[18] Under the vector representation of grammars, while parameter selection is independent, the fitness measure in the learning is defined on *whole grammars*. Do we need to assume that the learner is able to infer, backwards, what to do with individual parameters given their "collective" fitness as a grammar? In other words, should the update rules be modified as follows in parameter based learning?

(27) a. reward *all* the parameter values if the composite grammar succeeds.

b. punish *all* the parameter values if the composite grammar fails.

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, generally taken to be the movement of inflected verbs to Comp in matrix sentences, as in many Germanic languages. The language to be acquired is German, which has [+Wh] and [+V2]. Suppose that the parameter combination {+Wh, -V2} is chosen, and 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 and be punished. But how does the learner prevent the correct parameter value [+Wh] from punished? Similarly, the grammar {-Wh, +V2} will succeed at any declarative German sentence, and the wrong parameter value [-Wh], irrelevant to the input, may take a hitchhike and get rewarded.

A not unreasonable solution to this problem is to endow the learner with the ability to tease out the relevance of parameters when performing grammatical analysis. For example, one might assume that, based on the discourse context and the intonation, the learner can identify whether a sentence is declarative; if so, he will deduce that the Wh parameter plays no role in the analysis of the sentence and hence will not be affected. But this idea becomes less plausible if all 30-40 parameters are subject to such decision procedure every time a sentence is presented and analyzed.

A more plausible solution, it seems to me, should stick to the null hypothesis of a "dumb" learner and minimize the computational cost required in learning. Suppose param-

[18]I would like to thank Julie Legate, Steve Anderson, and Noam Chomsky for pointing out the seriousness of this problem.

eter learning is literally (27): let incorrect parameter values be rewarded as hitchhikers, correct parameter values be punished as accomplices, and hope, in the long run, the correct parameter values will prevail.

And this solution seems to work. Here I will give the result of a computer simulation; the reader is directed to Yang and Gutmann (in preparation) for a formal proof. Again, consider the learning of the two parameters [Wh] and [V2] in a German environment. The combinations of the two parameters give four grammars, which we can explicitly measure their 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 in [Spec,CP], which are only compatible with the [+V2] value. We then have a penalty probability chart:

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

We see that the two parameters, which wandered in fluctuation at earlier stages of learning, converge on the target values in the end.

Language acquisition in a parametric space gives rise to *partial grammars*. Since the successful acquisition of a grammar is accomplished only when all parameters are set correctly, children must go through stages in which some parameters are already in place while others are still in fluctuation. For example, an English child may have acquired that his language employs overt Wh movement, but has not conclusively determined whether it obligatorily uses overt subjects. Now what the child possesses are partial fragments of grammars that may correspond to any attested adult language. A number of empirical cases involving this scenario will be documented in chapter 3.

A most important consequence of learning by parameters lies in the dramatic (logarithmic) reduction of computational cost. It makes the competition model psychologically more

[19]This figure is based on English data: I am taking the liberty to extrapolate it to German.

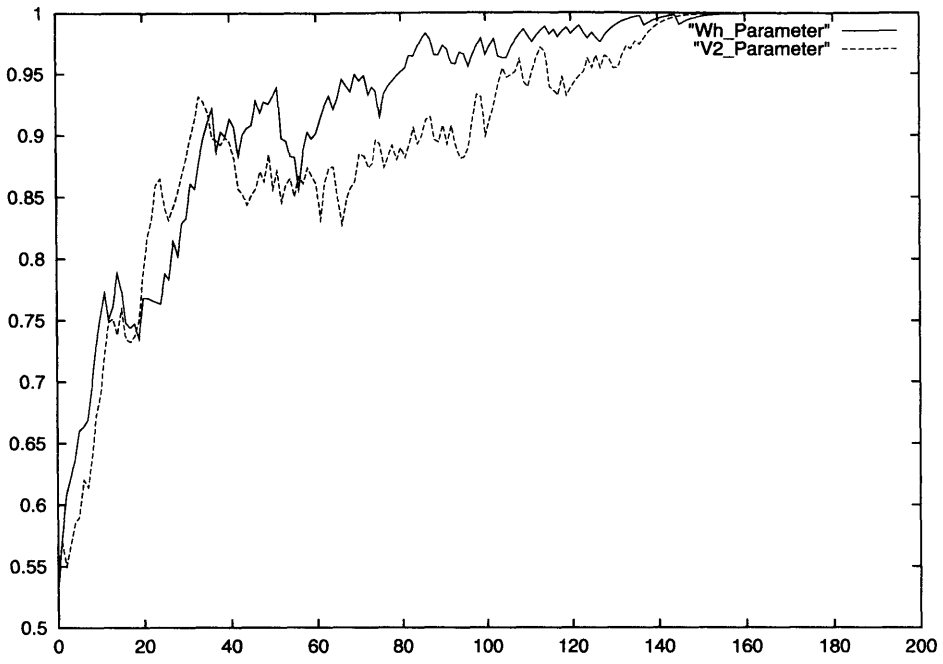

Figure 2-2: The independent learning of two parameters.

plausible, and in addition, gives a computational argument for the conception of UG as a parametric space.

2.4.2 Parameter Expression and Cues

While parametric learning discussed above resolves the general problem of parameter interference, there are specific cases where the problem does not even arise: some parameters can be learned entirely independently, irrespective of the values of other parameters.

Independent parameter learning was first studied by Dresher and Kaye (1990). They note that the parameters in a theory of 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. Clark's (1992) idea of *parameter expression* is similar. A sentence s expresses a parameter α if a grammar must have set α to some definite value in order to analyze s. In other words, expressive sentences can set the value of α correctly no matter what values other parameters are set to.

Dresher and Kaye (1990) propose that for each parameter, the learner is innately endowed with the knowledge of the cue associated with that parameter. Upon the presentation of a cue, the learner sets the value for the corresponding parameter, again in an all-or-nothing manner. (Fodor's unambiguous triggers (1998) can be seen as another sort of cues, which are determined by multiple parses.) In the variational model, the internalized cue-parameter association need not to be assumed. Rather, cues are interpreted as extensional expressions 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 cues, 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 cues have been encountered.

While cues that express parameters are not required by the current learning model (see section 2.3.3), they are quite useful in making developmental predictions. Suppose we have two parameters, each of which can be independently learned with cues. We can then estimate the frequencies of their respective cues, and make the prediction, based on (24), that the parameter with more frequent cues ought to be learned sooner than the one with less frequent cues. In Chapter 3, we will examine the acquisition of several parameters that can independently learned. One is the verb raising parameter that determines whether a finite verb raises to the Tense node: French sets this parameter to [+], and English, [−]. The [+] value for this parameter is associated with cues such as (63), where finite verbs precede negation/adverb. The other parameter is the optional subject parameter, for which the [−] value (English) is associated with expletive *there* sentences. Quantitative comparisons of parameter acquisition will be made there.

2.5 Related Approaches

The idea of language acquisition as grammar competition is not entirely novel, although it has never been worked out systematically or related to the developmental data in child language.

To the best of my knowledge, Jakobson (1941/1968) was the first to interpret "errors"

in child phonology as possible phonological forms in non-target languages. This position is echoed in Stampe (1979). Under the P&P framework, linguists have often viewed language acquisition as selecting a grammar out of all the possible human grammars (Piattelli-Palmarini 1989, Lightfoot 1991), although no concrete model of selection has been formalized and investigated. That children may have simultaneous access to multiple hypotheses has been suggested by Berwick and Weinberg (1984), Pinker (1984), and Pinker and Prince (1988), 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 approaches to the problems under study.

Most recently, Roeper (1999; cf., Yang 1999) has independently proposed that child language should be explained as a combination of multiple grammars which are 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, 1995): grammars with less complex structural representations are preferred.[20] 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. In Chapter 3, we give additional evidence for the reality of multiple grammars in child language development.

The genetic algorithm (GA) model of Clark (1992) bears closest resemblance to the present model. The GA model represents grammars as parameter vectors, which undergo reproduction via crossover (parts of a grammar are swapped/combined with others).[21] 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. It is clear that both the GA model and the variational model are explicitly built on the idea of

[20]Note that the proposed model is presented in a most generic way: all grammars are there to begin with, and input-grammar compatibility is the only criterion for rewarding/punishing grammars. This can of course be expanded to incorporate other possibilities, including the one suggested by Roeper. For instance, one can build in some appropriate prior bias in grammar evaluation that goes against complex grammars.

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

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 compute 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 seems too high 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, and by factoring the complexity of grammar competition into a parametric and combinatorial representation.

In the next Chapter 3, we pursue the condition of developmental compatibility and present evidence from children's syntax to support the variational model of acquisition, .

Chapter 3

Competing Grammars in Child Syntax

> Phylogenesis is the mechanical cause of ontogenesis. The connection between them is not of an external or superficial, but of a profound, intrinsic, and causal nature.
>
> Ernst Hackel (1874), quoted by
> Stephen Jay Gould in *Ontogeny and Phylogeny* (1977)

Hackel's "ontogeny recapitulates phylogeny", which has gone in and out of (and back in) fashion in biology, may well be vindicated in the ontogeny of human language, with a definite twist. If language is delimited in the finite space of Universal Grammar, its ontogeny may just recapitulate its scope and variations as the child gradually settles on one out of the many possibilities. This is exactly what the proposed variational model leads one to expect, and the present chapter documents quantitative evidence to demonstrate its validity.

If one surveys the field of language acquisition, we cannot fail to notice an unfortunate gap between formal models and developmental studies. We rarely find formal models attempting to explain directly patterns in children's language, or rigorous proposals of how the child attains and traverses the descriptive "developmental stages". The variational model, I believe, fills this gap.

The variational model makes two general predictions about child language development:

(28) a. Other things being equal, the rate of development is determined by the penalty probabilities of competing grammars; cf. (24).

b. As the target grammar gradually rises to dominance, the child entertains co-existing grammars, which ought to be reflected in the non-uniformity and inconsistency in its language.

What follows is a preliminary investigation of (28) through several case studies in children's syntactic development. These cases are selected because of the availability of carefully documented quantitative data in the literature; I am therefore indebted to the colleagues whose data the present chapter is based on. We will show that some interesting and important generalizations in the child language data cannot be revealed or explained unless a variational approach is assumed. Section 3.1 presents cross-linguistic longitudinal evidence in support of prediction (28a), drawing evidence from child French, English, and Dutch. In Section 3.2, we will develop a quantitative interpretation of the Argument from the Poverty of Stimulus presented in (1.1), in response to recent challenges by Sampson (1989) and Pullum (1995). Section 3.3 gives a systematic account of null subjects in child English, in comparison with child Chinese and Italian. Based on the children's null subject wh-questions and null object sentences, we show that English children have simultaneous access to both an obligatory subject grammar (the target) as well as an optional subject grammar, supporting prediction (28b). The case studies will be concluded with a "working manual" for acquisition studies in the variational framework.

3.1 The Time Courses of Three Parameters

Recall that in (24) (section 2.3.1), the penalty probability of the competitor grammar determine the rate of language development. We put the variational model to test by examining the acquisition of three parameters: that of French finite verb raising acquired early (Pierce 1989), that of English subject use, acquired relatively late (Valian 1991), and that of Dutch V2 parameter, also acquired late (Haegeman 1995). Following the discussion of parameter learning in section 2.4, we estimate the frequency of cues that unambiguously express the target value of each of three parameters under study.

3.1.1 Verb Raising and Subject Drop: the Baselines

Consider first the verb to Tense raising parameter, for which the [+] value is expressed by cues of the pattern V_{FIN} Neg/Adv.[1] A grammar with the [−] value for this parameter is incompatible with such sentences; when probabilistically selected by the learner, the grammar will be punished as a result. Based on the CHILDES corpus, we estimate that such cues constitute 7% of all French sentences children hear.[2] Since verb raising in French is an early acquisition (20th month; Pierce 1989), this suggests that 7% of unambiguous cues is a lower bound that suffices for an early acquisition: any aspect of grammar with at least 7% of cues should also be acquired very early.

We then have a direct explanation of the well-known observation that word order errors are "triflingly few" (Brown 1973:156) in children acquiring fixed word order languages. For example, English children rarely produce word orders other than SV/VO, nor do they fail to front wh-words in questions (Stromswold 1990). Observe that virtually all English sentences display rigid word order, e.g., verb almost always (immediately) precedes object. Also, wh-words are almost always fronted in questions, which, in our estimation, constitute roughly 1/3 of all sentences English children hear. These patterns give very high, far greater than 7%, rate of unambiguous cues, which suffices to drive out other word orders very early on.

From (28a) it also follows that if cues are rare in the input, the development of a grammar (or a parameter) will be relatively late. Consider then the acquisition of subject use in English. Following Hyams (1986), Jaeggli and Safir (1989), and many others, pure expletive (*there*) subject sentences such as (29) signify the obligatoriness of subject use in a language:

(29) a. There is a man in the room.

b. Are there toys on the floor?

Optional subject languages do not have to fill the subject position, and therefore do not need placeholder items such as *there*.[3] We estimate that expletive sentences constitute 1% of all

[1] 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, Shortest Move (Chomsky 1995b), prohibits such raising to "skip" the intermediate Tense node. Therefore, finite verbs followed by negation or adverbs in a language indicate that the verb raises *at least* to Tense in this language.

[2] I would like to thank Julie Legate for her assistant in this study.

[3] This does not mean that we are committed to the postulation of a parameter, [± pro-drop] or [± Null-

adult sentences to children based on the CHILDES database. Since subject use is acquired relatively late (36th month, Valian 1991), we may conclude that 1% of unambiguous evidence ought to result in a late acquisition. Similar to the case of French verb raising, we will use 1% as a baseline for *late* acquisition: if a parameter is expressed by 1% of all input, then its target value should be set relatively late; more specifically, comparable to child subject use in English.[4]

3.1.2 V1 Patterns in V2 Learners

Consider finally the acquisition of the V2 parameter in Dutch. As noted in (25), there appears to be no direct cues for the V2 parameter: the four competitor grammars together provide a complete covering of the V2 expressions. However, three competitors, namely, the English, Irish, and Hixkaryana type grammars, while compatible with SVO, XVSO, and OVS patterns respectively, nevertheless have very high penalty probabilities: 35.3%, 66%, and 98.7%, according to our corpus analysis. As a result, these grammars are eliminated quite early on. Figure (2.3.3), which is based on a computer simulation, is duplicated here.

The Arabic type grammar, a "type II" grammar in Greenberg's typology (1963), fares considerably better in the competition. By the virtue of allowing SVO and XVSO alternations, it is compatible with an overwhelming majority of V2 patterns (98.7% in all). However, it is not compatible with OVS sentences, which therefore are effectively unambiguous cue for the target V2 parameter *after* the other three competitors have been eliminated very rapidly. The rarity of OVS sentences (1.3%) implies that the V2 grammar is a relatively late acquisition, with an Arabic type non-V2 grammar in co-existence with the target V2 grammar for an extended period of time.

The presence of an Arabic type grammar allows verb-initial (V1) sentences, which are

Subject], which is in any case too crude to capture the distributional differences between two representative classes of optional subject grammars, the Italian type and the Chinese type. We only make the assumption that languages make a small number of choices with respect to the use of subject. We are however committed to what seems to be a correct generalization that the use of expletive subjects and the obligatoriness of subject are correlated – hence, something in UG must be responsible for this.

[4]Notice that while the acquisition of subject parameter is a late phenomenon (similarly, the acquisition of Dutch V2 to be discussed later), other parameters grammar can be and in fact are learned relatively early: the early mastery of word order and Wh-fronting discussed earlier are such examples. This means that at least the parameters relevant for these cases are set independently during learning, as suggested in section 2.3.3.

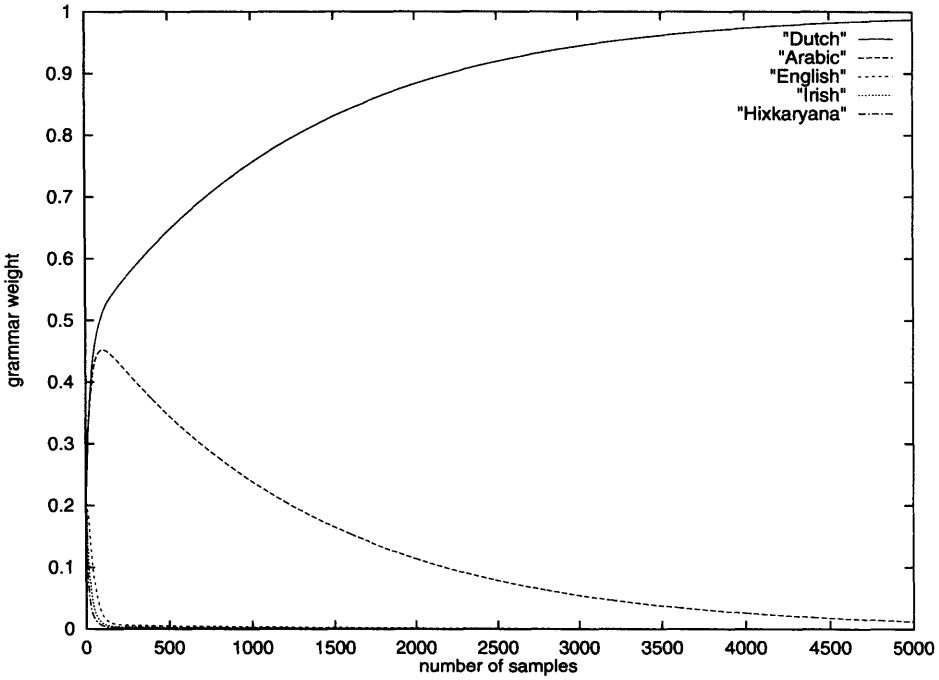

Figure 3-1: The emergence of the V2 grammar as a result of competition. Note the early elimination of the English, Irish, and Hixkaryana grammars.

ungrammatical in the target V2 grammar, but which will nevertheless constitute a significant portion of Dutch child language. This prediction is confirmed based on the statistics compiled by Haegeman (1995), one of the largest longitudinal studies in the acquisition of V2 languages. The longitudinal data are summarized in Haegeman's (1995) tables 6 and 7, combined in (3.1.2) below.

age	Preverbal Subject	Postverbal Subject	Overt Material Left of V
2;4	76	94	22
2;5	44	88	22
2;6	239	172	25
2;7	85	116	23
2;8	149	143	49
2;9	126	143	55
2;10	146	175	82
2;11	124	135	99
3;0	126	120	64
3;1	108	160	59

Table 3.1: Subjects and non-subject topics in Hein's Finite Clause

The total number of finite clauses is computed by summing column 2 and column 3 in (3.1.2). The number of V1 sentences is column 2 minus column 4 (XVS). It is important to note that all the sentences contain overt subjects, hence ruling the possibility that the superficial V1 patterns are due to subject drop, which Germanic children are known to do. The frequency of V1 sentences is shown in (3.1.2): We see that before 2;6, the child used V1 patterns in close to 50% of all sentences; see Wijnen (1999) for similar findings. This high level of V1 use contradicts the recent claims that the V2 parameter is set correctly very early on (Poeppel and Wexler 1993, Wexler 1998). It is also difficult to maintain performance-based approaches (Pinker 1984): attributing the 50% of V1 use to performance is as good as saying that children use the V2 grammar (adult-like competence) *randomly*.

As shown in Table (3.1.2), Hein's use of V1 sentences dropped to about 15% at 3;0. This can be interpreted as the target V2 grammar gradually wiping out the Arabic type grammar. Furthermore, because the frequency (1.3%) of Dutch OVS sentences is comparable to the frequency (1%) of English expletive sentences, we predict, on the basis of (24), that the

age	V1 sentences	all sentences	V1%
2;4	76	170	45%
2;5	66	132	50%
2;6	147	411	36%
2;7	93	201	46%
2;8	94	292	32%
2;9	98	269	36%
2;10	93	321	28%
2;11	36	259	14%
3;0	56	246	22%
3;1	101	268	37%[5]

Table 3.2: Hein's longitudinal V1 patterns

V2 parameter is successfully acquired roughly at the same time as when English children have adult-level subject use (36th month). If we use Brown's criterion that 90% correct use signals a successful acquisition, we may conclude that Dutch children have mastered V2 at 3;0 (perhaps a little later). There is evidence in the acquisition of German, a similar language, that children have acquired V2 by the 36-39th month (Clahsen 1986). Under the present model, it is not a coincidence that the timing of the acquisition of English subject use and that of Dutch/German V2 are comparable.

3.2 A Quantitative Argument from the Poverty of Stimulus

Based on the acquisition model and the findings in section (3.1), we can give a quantitative evaluation of the Argument from the Poverty of Stimulus (APS).[6]

Again, the issue is, why do human children unequivocally settle on the correct, structure dependent, rules for question formation, while the input evidence does not rule out the incorrect, structure independent, inductive generalization:

(30) a. front the first auxiliary verb in the sentence

b. front the auxiliary verb that is most closely follows a noun

[6]This section is based on the work of Julie Legate (1999), whom I am grateful for sharing her data and ideas with me, and of course, for many other things.

c. front the second word in the sentence

d. ...

for which the relevant evidence is many way ambiguous:

(31) a. Is Alex *e* singing a song?

b. Has Robin *e* finished reading?

Recently, the APS based on structure dependency in question formation has been challenged by Sampson (1989) and Pullum (1996). They claim that the learner is actually exposed to a sufficient amount of evidence to rule out the incorrect, structure independent, hypotheses. Here we will analyze Pullum's objections and show that they are not valid.

First, Pullum (implicitly) assumes that there is only *one* alternative hypothesis to be ruled out, namely, that of (30a), the inversion of the first auxiliary in the sentence. This assumption is incorrect: the learner in fact has to rule out *all*, in principle infinitely many, hypotheses compatible with (31), But for the sake of argument, suppose it were the case that the learner had only a binary choice to make, while keeping in mind that if the learner did not have prior knowledge of structure dependency, the effort it takes to rule out all possible hypotheses can only be harder than that to rule out (30a).

Second, Pullum notes, correctly, that auxiliary inversion in yes-no questions is not the only type of sentences that rules out (30a):[7]

(33) Is_1 [the boy who is]$_{\text{NP}}$ t_1 in the corner smiling?

Wh questions with an inverted auxiliary over a complex NP are also informative:

(34) How could_1 [anyone that was awake]$_{\text{NP}}$ t_1 not hear that?

[7]Less convincingly, Pullum argues, following Sampson (1989), that sentences like those in (32) disambiguate the correct rule from (30a) as well:

(32) If you don't need this, can I have it?

Pullum reasons that in (32), with the underlying representation of [*If you don't need this, I can have it*], the structure-independent rule (30a) would produce front either *don't* or *can*, producing erroneous output. This argument goes through only if it can be shown that the learner treats sentence boundaries equally as constituent boundaries in a single clause.

Pullum proceeds to count the frequency of sentences such as (3.2) and (3.2), using a Wall Street Journal corpus. He discovered that in the first 500 sentences he examined, 5, or 1%, are of the two types. Some examples are given below:

(35) a. How fundamental are the changes these events portend?

b. Is what I'm doing in the shareholders' best interest?

c. Is a young professional who lives in a bachelor condo as much a part of the middle class as a family in the suburbs?

d. Why did "The Cosby Show's" Lisa Bonet, who has a very strong screen presence, think that participating in a graphic sex scene would enhance her career as a legitimate actress?

Pullum then concludes that the APS is flawed, since the learner does have access to a non-trivial amount of disambiguating evidence.

This is a valid argument, only if situated in a *comparative* setting of language acquisition. That is, we need an independent yardstick to quantitatively relate the amount of relevant linguistic experience to the outcome of language acquisition — the variational model offers just.

First and foremost, we must take an *independent* case in acquisition, for which we have good knowledge of children's developmental time course, and for which we can also obtain a corpus count of the relevant evidence. The null subject phenomenon is a perfect example.

As reviewed, English children's subject use reaches adult level at around the 36th month (Valian 1991). This is comparable to the age of the children whose knowledge of structure dependence Crain and Nakayama (1987) tested: the youngest group was at 3;2. In both cases, the learner will make a binary choice: Valian's children have to determine whether the language uses overt subjects, and Crain and Nakayama's children have to rule out that the language uses the structure-independent rule, "invert first auxiliary". Under the present model, in fact, under *any* reasonable quantitative model, comparability in the time courses of two acquisitions must entail comparability in the frequencies of their respective evidence. If English subject use is gradually learned on the basis of "*there*" expletive sentences, which represent roughly 1% of all sentences, then one would expect sentences of the type (3.2) and

(3.2), which supposedly serve to establish structure dependence, must also be close to 1% in the input data.

Which takes us to a second problem in Pullum's argument: we must start with *realistic* corpora of children's linguistic input. The Wall Street Journal hardly fits the bill, a point that Pullum acknowledges. Based on the CHILDES transcripts of the adult utterances that an American child Adam was exposed to, Legate (1999) finds the following:

(36) In a total of 20,372 sentences, 8,889 were questions, of which

a. No yes-no question of the type (3.2) was found.

b. Four Wh questions of the type (3.2) was found:[8]

i. Where's the part that goes in between?

ii. What is the music it's playing?

iii. What's that you're drawing?

iv. What was that game you were playing that I heard downstairs?

Not only is this far below the 1% that we estimate would be required for the child to learn the correct rule by the 36th month, but also low enough to be considered negligible, or to be available for *every* human child. We hence conclude that the original APS stands unchallenged: the knowledge of structure dependence in syntax, as far as we can test quantitatively and comparatively, is available to children in the absence of experience.

3.3 The Nature of Null Subjects in Children

We now turn to a detailed analysis of null subjects (NS) in English children in comparison to Chinese and Italian children. We begin with a typology of subject use across languages to establish the nature of the grammars that compete during acquisition.

To recover the referential content of a null subject, optional subject grammars employ one of the two (almost inevitable) strategies (Huang 1984). In languages like Italian and

[8]Of these, it is not even clear whether the equative sentences (36b-iii) and (36b-iv) should count as evidence against (30a). The child might analyze them with the wh-word in the subject position and the complex NP is the object position (although this is arguably not the analysis ascribed to these questions in adult grammar).

Spanish, a null subject is identified via unambiguous agreement (number, person, gender) morphology on the verb; in languages like Chinese and Japanese, a null subject is identified via linking to a discourse topic, which serves as its antecedent. Because of the differences in the identification mechanism, Chinese and Italian show different distributions of null subjects.

First, Italian does not allow (arbitrary) null objects (NO) (Rizzi 1986). In contrast, Chinese does freely allow NO (Huang 1984), which, like null subjects, can be recovered by linking the empty pronominal to a discourse topic:

(37) TOPIC_1 [Zhangsan kanjian-le e_1]. (e_1 = him)
TOPIC_1 [Zhangsan saw-ASP him_1].
'Zhangsan saw him.'

However, Chinese NS is more restrictive than Italian. When a topic phrase (Top) is fronted, subject drop in Chinese is grammatical only if Top is not a possible antecedent for the null subject, for otherwise the linking to discourse topic is disrupted. More specifically, Chinese NS is possible (38a) when Top is an adjunct, which can never be the antecedent of a dropped subject, and not possible (38b) when TOP is an argument (object).

(38) a. Zai gongyuan-li_2, [e_1 t_2 da-le ren]. (e_1 = John)
In park-LOC, [e_1 t_2 beat-ASP people].
'It is in the park (but not at school) that John beat people up.'

b. *Sue_2, [e_1 xihuan t_2]. (e_1 = John)
Sue_2, [e_1 likes t_2].
'It is Sue (but not Mary) that John likes.'

Italian identifies null subjects through Agreement morphology, and does not have the restrictions on subject drop seen above in Chinese. Subjects can be dropped freely in nominal and non-nominal Wh questions,[9] as shown below:

(39) a. Chi_2 e_1 ha baciato t_2?
Who_2 has(3SGM) kissed t_2?
'Who has he kissed?'

[9]Following Chomsky (1977) and many others, we assume the generalization that topicalization and Wh-movement are essentially the same process (movement to [Spec,CP]), for they share many syntactic and semantic properties. Since Chinese cannot front Wh phrases (in questions or any other constructions), only topicalization data can be given in (38).

b. Chi_3 e_1 credi che e_2 ami t_3?
Who_3 e_1 think(2SG) that e_2 loves(3SGF) t_3?
'Who do you think she loves?'

c. Dove_2 hai e_1 visto Maria t_2?
Where_2 have(2SG) e_1 seen Maria t_2?
'Where have you seen Maria?'

The differences between Chinese, English, Italian subject use are summarized below:

(40) Chinese: object drop, no subject drop with argument topicalization
English: no object drop, obligatory subject, use of expletive *there*
Italian: no object drop, unrestricted subject drop, rich Agreement morphology

We shall see how such differences play out their roles in child language acquisition, disambiguating these grammars from one another. In addition, we shall see how these differences are recapitulated in (English) children's acquisition of subject use. I will again stress that the learner does not actively search for the patterns in (39) to identify their target grammar, as in a cue-based learning model. Rather, the grammars are probabilistically selected to analyze incoming sentences, and they will face different outcomes in different linguistic environments. For example, both English and Italian grammars will be punished in a Chinese environment when a null object sentence is encountered. Only the target grammar wins out in the end.

3.3.1 The Early Acquisition of Chinese and Italian Subject Drop

Here we study the acquisition of subject use in Chinese and Italian children; we turn to English children in 3.3.2. Throughout our discussion, when we refer to a particular grammar (for example, Chinese), we mean the property of subject use in that *type* of grammar (discourse-based subject drop).

Consider first how a Chinese child rules out English and Italian grammars. Here, null object sentences like (37) are unambiguous cues for the Chinese, discourse-based subject drop grammar. A study by Wang *et al.* (1992) shows that Chinese adults use a fair amount of object drop sentences in speech to children (11.6%, computed from their Appendix B) as well as among themselves (18%, computed from their Appendix D). In section (3.1),

we have empirically established that 7% of unambiguous evidence suffices for an very early acquisition, as in the mastery of finite verb raising by French children (Pierce 1989). We thus predict that from very early on, Chinese children have eliminated English and Italian grammars, and converged on the remaining grammar, the target.

This prediction is borne out: Wang *et al.* (1992) find that Chinese children's use of subject drop and object drop is not only fairly constant for all age groups, but also comparable to that of adults. The results are summarized in Table (3.3.1), based on their Appendices C and D:

Age	Subject Drop (%)	Object Drop (%)
2	55.728	20.192
3	45.650	21.376
4	38.252	26.031
overall	46.543	22.533
adults	45.852	18.000

Table 3.3: Chinese adult and child pronominal drop

Let's now turn to Italian children. Recall from earlier discussion that Chinese does not allow subject drop when an argument assumes the topic position (38b), and Italian does (with a fronted argument Wh phrase. This means that every subjectless question with an argument (specifically, object) wh-question punishes a Chinese grammar, and of course an English grammar as well.

It is known that in adult speech to children, approximately 70% of all utterances have dropped subjects (Bates 1976, Caselli *et al.* 1995). We also know that Wh questions are one of most frequent constructions children are exposed to. We estimate that about 15% of all sentences are object questions involving empty subjects – the lower bound of 7% is met to warrant an early acquisition. This prediction is confirmed by Valian's findings (1991): at both of the developmental stages investigated (1;6 to 1;10 and 2;0 to 2;5), Italian children drop subjects in about 70% of sentences, comparable to the figures in adult speech reported in Bates (1976).

3.3.2 Why English Kids (Sometimes) Use Chinese

Finally, we consider how English children come to know that their language is an obligatory subject grammar, ruling out the Chinese and Italian grammars that are also made available by UG.

We first claim that the Italian grammar can very rapidly be eliminated by English children on the basis of their knowledge of agreement morphology. There is strong evidence that young children's agreement morphology is virtually perfect. Phillips (1995:327), reviewing a number of cross-linguistic studies, observes that "in languages with overt agreement morphology, children almost always use the agreement morphemes appropriate to the argument being agreed with". For example, Guasti (1994) found that three young Italian children use agreement morphology correctly in more than 95% of all contexts. Clahsen and Penke (1992) had similar findings in a German child during the period of 1;7 to 2;8: the correct use of agreement with the affixes *-st* (2nd singular) and *-t* (3rd singular) is consistently above 90%.[10] Children's near-perfect knowledge of agreement morphology plays an important role in grammar competition. Specifically, if an Italian grammar is chosen to analyze English input, the lack of unambiguous agreement in English causes the Italian grammar to fail and be punished as a result.

The Chinese grammar is more difficult to rule out. Chinese employs discourse linking as the mechanism for null subject identification; morphology provides no useful information. The only evidence against the Chinese grammar is expletive *there* sentences, which constitute only 1% of all input sentences. Hence, with respect to subject use, we predict that English children ought to entertain an English grammar in co-existence with a Chinese grammar for an extended period of time.

The claim of grammar co-existence attributes English child NS to the presence of the Chinese grammar, which is probabilistically accessed. This directly explains the fact that English children use a non-trivial amount of NS, but at a lower rate (30%) than Chinese children (46.5%) (Wang *et al.* 1992). We also predict that child English ought to contain

[10] When children do deviate from adult forms, their morphological errors are overwhelmingly of omission, the use of a default form, rather than substitution, the use of an incorrect form. In Chapter 4, we will see that this pattern is strikingly similar to that of the extensively studied English verb past tense morphology; both of them, I will argue, follow a general model of morphological acquisition proposed there.

a certain amount of null objects (NO), grammatical in Chinese. Such an account of NO does not appeal to additional assumptions such as performance factors (Bloom 1993). Furthermore, the presence of the Chinese grammar entails that the distributional patterns of English child NS ought to show characteristics of a Chinese grammar. To demonstrate this, we make two quantitative predictions that are borne out below.

First, recall that characteristic of a Chinese type grammar, NS is only possible in adjunct topicalizations (38a), but not in argument topicalizations (38b). Since we attribute English child NS to the presence of a Chinese grammar, we predict that NS will be possible in adjunct questions but not possible in argument (object) questions.[11] This prediction is confirmed as follows. During the NS stage of Adam (CHILDES: files 1-20), we found an almost categorical asymmetry of NS in adjunct and argument questions:

(41) a. 95% (114/120) of Wh-questions with NS are adjunct (*how, where*) questions.

b. 97.2% (209/215) of object questions (*who, what*) contain subjects.

The second prediction concerns the relative frequencies of NS and NO. Since both NS and NO are attributed to the Chinese grammar, we predict the relative ratio of NS/OS to hold fairly constant across English and Chinese children in a same age group. This prediction is made as follows. Suppose that for Chinese children, NS ratio is s and NO ratio is o, and that for English children, NS ratio is s' and NO ratio is o'. Suppose further that, during the NS stage, English children access the Chinese grammar with the probability p, which leads to the NS and OS patterns in production. Recall ((3.3.1)) that Chinese children learn their grammar very early, showing adult-like performance; they use the Chinese grammar 100% of the time. Now if we scale up p to 100%, that is, English children were to use the Chinese grammar *mono-lingually*, we expect that their NS and OS ratios to be identical to those for Chinese children.[12] That is, $s' = sp$ and $o' = op$, which implies $s'/o' = s/o$.

The confirmation for this prediction is shown in Table 3.3.2, based on the statistics reported in Wang *et al.* (1992):[13]

[11] The fronting of the Wh word in question formation, of course, is an early acquisition, as noted in section 3.1. Again, the parameter for Wh-fronting and the subject parameter are set independently.

[12] Assuming, without evidence to the contrary, that English and Chinese children are equally likely to encounter discourse situations in which NS and OS would be employed.

[13] We have used the statistics for American children between 2;0 and 3;0 (the NS stage, Valian (1991)).

Language	Null Subject (NS) %	Null Object (NO) %	NO/NS %
Chinese	55.728	20.192	36.233
English	25.885	8.308	32.096

Table 3.4: Chinese and English child subject and object drop.

The quantitative predictions reported here, including the categorical asymmetry in argument and adjunct questions and the relative ratio of NS/NO, fall out naturally under the variational model of grammar competition, which explicitly appeals to the syntactic properties of competing UG grammars given by theories of adult linguistic competence. They cannot be made under performance-based theories (Bloom 1990, 1993, Gerken 1991, Valian 1991) that assume English children have an adult-like a single, obligatory subject grammar and that null subjects result from performance factors that perturb the use of their grammar.[14] The recent OI infinitive based approach to null subject (Sano and Hyams 1994, Hyams 1996, Wexler 1998, among others). which holds that null subjects are licensed by non-finite root verbs,[15] also cannot explain these generalizations.

3.4 Summary

We summarize the key features and results of the variational model as applied to syntactic acquisition:

(42) a. Language acquisition can be modeled as a selectionist process in which variant grammars compete to match linguistic evidence.

b. Under the condition of explanatory continuity, the irregularity in child language

The NS ratio for American children is after the adjustment to rule out NS sentences that would have been acceptable in adult English.

[14]There is a rather large body of literature against the performance based approach to NS; see, e.g., Hyams and Wexler (1993), Roeper and Rohrbarher (1994), Bromberg and Wexler (1995), Waller (1997), among others.

[15]Note that this claim predicts that the OI stage and the NS stage should end at roughly the same time. There is *prima facie* evidence against this prediction. For example, the OI stage for a Dutch child Hein (Haegeman 1995: Table 4) essentially ended at 3;0 and 3;1, when his RI usage dropped to 4% and 6%. However, at 3;0 and 3;1, there were still 30% and 31% of NS sentences. See Phillips (1995) for additional evidence that the correlation between OI and NS is weak.

and the gradualness of language development can be attributed to a probabilistic combination of multiple grammars, rather than an imperfect grasp/use of a single grammar.

c. Formal sufficiency and developmental compatibility are met in a unified model, for which the course of acquisition is determined by the relative compatibilities of the grammars with input data; such compatibilities, expressed in penalty probabilities, are quantifiable and empirically testable.

The variational theory offers a new interpretation of all aspects of children's language. The first step is the observation of non-uniformity in children's language, specifically, the deviation from the adult grammar they are acquiring. Second, we try to identify the grammars, which are not what the learner is exposed to but nevertheless are options allowed by UG (and possibly attested in the world of existing languages), and which, *collectively* with the target grammar, give a complete coverage of children's language. Third, we associate each of the competing grammars with its corresponding disconfirming evidence in the linguistic environment, that is, input patterns that they are incompatible with.[16] Finally, we use naturalistic adult-to-child linguistic database such as CHILDES to access the penalty probabilities of the competing grammars, which are then used to make quantitative predictions as in section 3.1 and 3.2. Furthermore, the idiosyncratic properties of co-existing competing grammars will be recapitulated in children's language, as demonstrated in section 3.1.2 and 3.2. In future work, this procedure will be systematically applied to a wide range of topics in child language, along the lines of research sketched out in this chapter.

[16]It is clear that both step two and three are guided by linguistic theories and typology; more on this in Chapter 6.

Chapter 4

Words, Rules, and Competitions

Fuck these irregular verbs.

Quang Phuc Dong (1968/1971)

The acquisition of the past tense in English has generated much interest and controversy in cognitive science. This is unfortunate. The problem of past tense, particularly in English, notorious for its impoverished morphology, is a fairly marginal problem in linguistics: placing it in the center of attention does no justice to the intricacy of language and the enormous body of work dedicated to the cross-linguistic study of language acquisition; see Yang (2000) for general discussion. This is not to say the problem of English past tense is trivial or uninteresting. As we shall see, despite the enthusiasm on both sides of the past tense debate, there are still very important patterns in the past tense acquisition data unnoted and unexplained in previous works. We show that the variational learning model, instantiated here as a competition among morphophonological rules (rather than grammars/parameters, as in the case of syntactic acquisition), provides a new understanding of how the mind learns and organizes verbal past tense.

4.1 Background

Our problem primarily concerns the nature of three systematic patterns in children's use of past tense. First, it has been known since Berko's (1958) classic work that in general,

children inflect novel verbs with the -d suffix as in *rick-ricked*. Second, young children sometimes *overregularize*: for example, they produce *take-taked* instead of *take-took*, where the suffix -d for regular verbs is used for an irregular verb. On average, overregularization occurs in about 10% of all instances of irregular verbs, according to the most extensive study of past tense acquisition (Marcus, Pinker, Ullman, Hollander, Rosen, and Xu 1992). Third, errors such as *bring-brang* and *wipe-wope*, where children misapply and overapply irregular past tense forms, are exceeding rare – about 0.2% of all instances of irregular verb uses (Xu and Pinker 1995).

One leading approach to the problem of past tense, following the influential work of Rumelhart and McClelland (1986), claims that the systematic patterns noted above emerge from the statistical properties of the input data presented to connectionist networks. A number of problems with the connectionist approach have been identified (e.g., Fodor and Pylyshen 1988, Lachter and Bever 1988, Pinker and Prince 1988, Marcus et al. 1992; among numerous others). To give just one example (from Prasada and Pinker 1993), when novel verbs such as *slace* and *smeeb* are presented to a trained connectionist model, *fraced* and *imin* are produced as their respective past tense forms – a behavior clearly incompatible with human performance.

In this chapter, we will critically assess another leading approach to the problem of past tense, the *Words and Rule* (WR) model developed by Pinker and associates; see Pinker (1995, 1999) for a summary. The WR model claims that the computational system for past tense consists of two modules. In the *rule* module, following the tradition of generative linguistics, regular verbs are inflected by making use of a *default* morphophonological rule, which adds -d to the root (stem). This explains the productivity of -d suffixation to novel verbs. Equally important to the WR model is the Blocking Principle (Kiparsky 1973), a traditional linguistic idea dating back to Pāṇini. In past tense formation, the Blocking Principle has the effect of forcing the use of a more specific form over a more general form: for example, *sang* is a more specific realization of the past tense of *sing* than *singed*, and is therefore used. Irregular verbs are learned in the *word* module, by memorization of the pairing between the stem and the past tense form – much like associationist learning in connectionist models. The strength of association is determined by the frequencies of regular verbs that children hear;

thus, memorization of irregular verbs takes time and experience to be perfected. When the child's memory for an irregular form fails, the default -d form is used. This accounts for the second salient pattern of past tense acquisition: overregularization errors in child language.

But more importantly, we will advance an alternative approach to the acquisition of past tense, dubbed the *Rules and Competition* (RC) model. The RC model treats both irregular and regular verbs within a single component of the cognitive system: generative morphophonology. Like the WR model, we assume the presence of a default rule, which attaches the -d suffix to the stem and in principle applies to all verbs. In contrast to the WR model, we claim that past tense of irregular verbs are also formed by morphophonological rules.

The RC model derives from the variational approach to language acquisition developed in preceding chapters. The variational approach holds that the child learner's hypothesis space consists of multiple competing hypotheses. These hypotheses are associated with weights, and it is the weights, or the distribution of the grammars, that change in adaptation to the linguistic evidence in the environment. For the problem of past tense, the hypothesis space for each irregular verb x includes a rule $R_{\mathcal{S}}$, defined over a verb class $\mathcal{S}$ of verbs of which x is a member. For example, the rule [-t suffixation & Vowel Shortening] applies to irregular verbs such as *lose*, *deal*, and *dream*. The acquisition of x involves a process of competition between $R_{\mathcal{S}}$ and the default -d rule, which in principle could apply to all verbs, regular and irregular. The child learns from experience that for irregular verbs, irregular rules must apply, and the default -d rule must not. Before learning is complete, the default rule will be probabilistically accessed, which leads to overregularization errors.

Section 4.2 presents the RC model in detail, including a description of the past tense formation rules in the computational system and a learning algorithm that specifies how such rules compete. We will also give a learning-theoretic interpretation and revision of the Blocking Principle that underlies the WR model as well as much of generative morphophonology. Section 4.3 compares the WR and RC models, based on the child production data reported in Marcus et al. (1992). Specifically, we show that the correct usage of an irregular verb is strongly correlated with the weight of its corresponding morphophonological rule, which explains a number of class-based patterns in the acquisition of irregular

verbs. Such patterns receive no explanation under the the WR model, to the extent the WR model has anything explicit to say about these patterns. Section (4.4) takes on the proposal of pairing stem and past tense with analogy or phonological similarity in the WR model, which one might consider a partial remedy for the problems revealed in section (4.3). We show that insofar as learning by analogy is made concrete, it cannot capture the patterns uncovered in section (4.3). Section (4.5) gives a critical review of ten arguments that have been given in support of the WR model (Pinker 1995). We show that each of them is either empirically flawed or can be accommodated equally well in the RC model.

4.2 A Model of Rule Competition

A central question for a theory of past tense formation, and consequently, for its acquisition, is the following: should the -d rule be considered together with the inflection of the irregular as an integrated computational system, or should they be treated by different modules of cognition? The approach advocated here is rooted in the first tradition, along the lines pursued in Chomsky and Halle (1968), Halle and Mohanan (1985), and the present-day Distributed Morphology (Halle & Marantz 1993).

> These rules of verbal inflection constitute a continuum of productivity and generality that extends from affixation of the *-ed* suffix in *decide-decided* to total suppletion in *go-went*. ... In an intermediate class of cases exemplified by verbs like *sing-sang* or *bind-bound* the changes affect only a specific number of verbs. To deal with such cases, the grammar will not contain a plethora of statements such as "the past tense of *sing* is *sang*, the past tense of *bind* is *bound*," etc. Rather, it will contain a few rules, each of which determines the stem vowels of a list of verbs specifically marked to undergo the rule in question. (Halle & Mohanan 1985, p. 104)

This approach differs from the WR model, in which irregular verbs are individually memorized, to the effect of having "a plethora of statements".

4.2.1 A Simple Learning Task

Before diving into the details of our model, let's consider a simple learning task, which may help the reader understand the problem of past tense acquisition at a conceptual level. Suppose one is asked to memorize the following sequences of pairs of numbers (x, y):

(43) (2,4), (3,4), (4,8), (5, 10), (6,7), (7, 8), (8, 16)

Obviously, one strategy to do this is to memorize all the pairs in (43) *by rote*. The learner will store in its memory a list of pairs, *as is*: $(2,4),(3,4),...$ However, there is another strategy, which, when available, most of us are likely to employ. Notice that (43) contains two regularities between the two paired numbers (x, y) that can be formulated as two rules: $y = x + 1$ for {3, 6, 7} and $y = 2x$ for {2, 4, 5, 8}. In the memory of a learner that employs the second strategy, a list of x's will be associated with the rule that generates the corresponding y's:

(44) $\{3,6,7\} \mapsto R_{x+1}$

$\{2,4,5,8\} \mapsto R_{2x}$

We liken the acquisition of irregular verbs to the number sequence learning task described here. The WR model employs the first strategy: irregular verbs are memorized by rote as associated pairs such as *feed-fed*, *bring-brought*, *shoot-shot*, *think-thought*, etc. The RC model, based on a system of generative morphophonological rules, employs the second strategy such that irregular verbs are organized by rules that apply to a class of individuals:

(45) {feed, shoot, ... } $\mapsto R_{\text{Vowel Shortening}}$

{bring, think, ... } $\mapsto R_{\text{Rime}\rightarrow u}$

In an information-theoretic sense, the rule-based strategy, which allows a more "compact" description of the data, is the more efficient one.[1] Furthermore, there is reason to believe that the rule-based strategy is preferred when verbs (instead of numbers) are involved. While the number pairing rules can be entirely arbitrary and mentally taxing, the rules for irregular

[1]While the saving achieved by the use of rule may not be significant for English irregular verbs — there are only some 150 in all — it becomes dramatic when we move to other languages. This, along with the issue of irregular morphophonology in other languages, will be discussed in section 4.4.

verbs are *not*. The sound changing rules are often well-motivated phonological processes that are abundantly attested in the language. For example, the rule of Vowel Shortening[2] for verbs such *lose, feel, say*, etc., which shortens the long vowel in closed syllables followed by -d, -ø, and -t suffixes, is attested in many other suffixation processes in English (Myers 1987). Therefore, such rules are frequently encountered by and naturally available to the learner.

With this conceptual background, let's move on to the RC model. In what follows, we will describe the properties of the morphophonological rules for past tense, and how they compete in the process of learning.

4.2.2 Rules

The past tense rules in English fall into two broad dimensions: *suffixation* and *readjustment* (Halle 1990). Suffixation attaches one of the three past tense suffixes, -d, -t, and -ø (null morpheme), to the verb stem. Readjustment rules, mostly vowel changing processes, further alter the phonological structure of the stem.

We assume, along with the WR model, that as part of innate Universal Grammar, the child language learner is equipped with the knowledge of a *default rule*, which applies when "all others fail". The default rule for English verb past tense is given in (46):

(46) *The default -d rule*:

$x \xrightarrow{-d} x+$ -d

Irregular verbs fall into a number of *classes* as they undergo identical or similar suffixation and readjustment processes. Thus, verbs in a class are organized by a shared rule, which uniformly applies. Such a rule is schematically shown in (47), while the complete rule system for English past tense is given in Appendix A.

(47) *Rule $R_{\mathcal{S}}$ for verb class $\mathcal{S}$*:

$x \xrightarrow{R_{\mathcal{S}}} y$ where $x \in \mathcal{S} = \{x_1, x_2, x_3, \ldots\}$

[2]The term "Vowel Shortening" is perhaps a a misnomer. The change in the quality of the vowel actually involves the shortening *and* the lowering. While keeping this technical issue in mind, we will nevertheless continue to call such processes vowel lowering.

For example, the verb class consisting of *lose, deal, feel, keep, sleep*, etc. employs R = [-t & Vowel Shortening] to form past tense. It is important to realize that suffixation and readjustment rules are generally independent of each other, and are in fact acquired separately. For example, the suffixes in derivational morphology such as *ity, -al*, and *-tion* must be acquired separately, but they all interact with Vowel Shortening, a readjustment rule that applies to closed syllables under many kinds of suffixation, as shown by Myers (1987):

(48) Vowel Shortening in Suffixation:

[ay]-[I]: div**i**ne-div**i**nity

[i]-[ɛ]: d**ee**p-d**e**pth

[e]-[æ]: n**a**tion-n**a**tional

[o]-[ɑ]: c**o**ne-c**o**nic

[u]-[ʌ]: ded**u**ce-ded**u**ction

It is natural to expect that pervasive rules like Vowel Shortening can be readily built in the speaker's phonology to serve to define verb classes.[3]

Now the conceptual similarities and differences between the WR model and the RC model ought to be clear. It is not the case that the role of memory is completely dispensed with in the RC model. Every theory must have some memory component for verbs: irregularities, be definition, can not be predicted from either its sound or meaning, and hence must be recognized and somehow memorized. The difference lies in *how irregular verbs are organized*. In the WR model, irregular verbs and their past tense forms are stored as simple associated pairs, and learning is a matter of strengthening their connections. In the RC model, irregular verbs and their past tense forms are related by phonological rules (suffixation and readjustment), as schematically shown in the figure below:

One a rule system such as (47) is situated in a model of learning, a number of important questions immediately arise:

[3]It is worth nothing that some irregular verbs are conventionally grouped into vowel shifting classes, e.g. ablaut and umlaut, that are not as homogeneous as the Vowel Shortening class. ablaut and umlaut only designate the *direction* of vowel shifting, e.g. front → back, but leaves other articulatory positions, e.g. [± high/low], unspecified. Hence, further refinement is required within these heterogeneous classes (see Appendix A). We will return to the issue of class homogeneity in section 4.4.

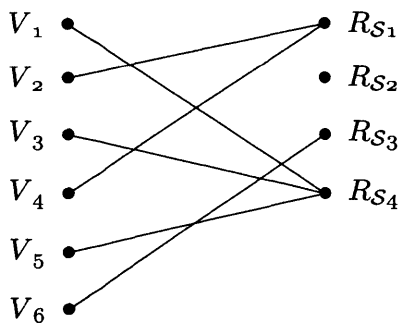

Figure 4-1: Each verb is associated with a rule-defined class.

(49) a. Where do rules such as suffixation and readjustment come from?

b. How does the learner determine the default rule (-d)?

c. How does the learner know which class a verb belongs to?

d. How do the rules apply to generate past tense verbs?

We postpone (49c) and (49d) to section (4.2.3). Consider (49a). First, we assume that UG, which encodes constraints on and variabilities in human language, plays a crucial role in the acquisition of morphophonology. Suppose that UG makes suffixation a possibility to express tense, person, gender, and other features. We assume that the child is able to extract -t, -ø, and -d suffixes from past tense verbs. We also assume that the child can arrive at the appropriate sound changing readjustment rules that relate the stem to the derived past tense form, as there is evidence that the child learner is sensitive to the phonotactics in its native language from very early on in life (Mehler, Jusczyk, Lambertz, Halstead, Bertoncini, and Amiel-Tison 1988, Jusczyk, Kelmer Nelson, Hirsh-Pasek, Kennedy, Woodward, and Piwoz 1992). In any case, the very fact that young children use past tense verbs correctly at a very high rate (about 90%, Marcus et al., 1992) and very rarely misuse irregular rules (about 0.2%, Xu & Pinker, 1995) is consistent with the assumption that children's knowledge of irregular rules is almost perfect.[4]

Consider then (49b), a problem that the WR model also has to address. Following

[4]Although young children have difficulty in producing multisyllabic words (e.g., Kehoe and Stoel-Gammon 1997), their performance on past tense is not affected: virtually all irregular verbs are monosyllabic, with a few exceptions of compounding origins, e.g. *forget, forseek, partake*, etc.

Pinker (1999, Chapter 8), children may derive the -d suffixation as the default past tense rule, if they are sensitive to the *type* frequency of regular verbs relative to all verbs they hear. Although the *token* frequency of irregular verbs is far greater than regular verbs, 86 percent of the most commonly used 1,000 English verbs are regular (Francis & Kucĕra 1982). In other words, the -d suffixation rule captures by far the largest *class* of verbs. Certain complications arise for the acquisition of participles in a language like German. From the top 1,000 most frequent verbs, the default suffix -t and the strong verb suffix -n are used at similar frequencies (about 45%), which appears to pose a problem for identifying the default rule based on type frequency. However, as noted by Pinker (1999: 217-221), citing Marcus, Brinkmann, Clahsen, Wiese, and Pinker (1995), the frequency of the -t suffix increases dramatically if more verbs are included. It increases further, if one collapses verbs that differ only in prefixes, e.g. *aukommen, aufkommen*, and *bekommen* into a single type. After such adjustment, 86% of German verbs have the default suffix -t. For a computer implementation of how default rules are acquired, with revisions to Pinker's proposals, see Molnar (forthcoming).

In the WR model, it is assumed that the default -d rule is not available until a little before the child's 3rd birthday (Pinker 1995). In section (4.5.3), we give empirical evidence against this view. Instead, we assume that that -d suffixation is recognized by the child as the default rule and is available from fairly early on, along with other suffixation and readjustment rules.

4.2.3 Rule Competition

Class Membership

We now address the question (49c), how children learn the class membership of irregular verbs. First, we assume, uncontroversially, that children are able to identify root and past tense pairs: for example, when *sat* is heard, the learner is able to deduce from the meaning of the sentence that *sat* is the past tense realization of the root *sit*. Once the root is extracted, the learner can proceed to associate it with the appropriate rule-based class.

It is logically possible that children may put a verb into a wrong class. However, there is strong empirical evidence against this possibility. Overwhelmingly, past tense errors are

overregularizations errors, which on average occur in about 10% of all instances of irregular verbs (Marcus et al., 1992). In contrast, misapplication and overapplication of irregular rules, such as *bring-brang*, *trick-truck*, *wipe-wope*, dubbed "weird past tense forms" by Xu and Pinker (1995) where a regular verb is irregularized or an irregular verb is irregularized incorrectly, are exceedingly rare – about 0.2 percent (ibid).[5] The rarity of weird past tense forms suggests that the child is *conservative* in learning verb class membership: without seeing evidence that a verb is irregular, the child assumes that it is regular, instead of postulating class membership arbitrarily.

Some notations. Write $P(x \in \mathcal{S})$ for the probability that the learner correctly places x into the verb class $\mathcal{S}$. Also, write f_x for the frequency of x in past tense form in the input, and $f_{\mathcal{S}} = \Sigma_{x \in \mathcal{S}} f_x$ for the frequency of a verb *class*, which in the sum of the frequencies of all its members. These frequencies, of course, can be estimated from adult-to-child corpora such as the CHILDES database (MacWhinney and Snow 1990).

Learning by Competition

We now turn the central component of the RC model: how rules apply to generate past tense verbs, and consequently, how they model the learning behaviors in children's use of irregular verbs.

The most important feature of the RC model is, *rule application is not absolute.* That is, every irregular rule $R_{\mathcal{S}}$, which applies to the verb class $\mathcal{S}$, is associated with a weight (or probability) $P_{\mathcal{S}}$. For example, when the child tries to inflect *sing*, the irregular rule [-ø & umlaut], which would produce *sang*, applies with a probability that might be less than 1. This follows if learning is *gradual*: it does not alter its grammar too radically upon the presentation of a single piece of linguistic evidence. The -d rule applies as the default, with probability $1 - P_{\mathcal{S}}$, when the irregular rule $R_{\mathcal{S}}$ fails to apply.[6]

[5] See Clahsen and Rothweiler (1993) for similar findings in German acquisition.

[6] The present model should not be confused with a suggestion in Pinker and Prince (1988), which has an altogether different conception of "competition". Pinker and Prince suggest, much like the present model, that irregular verbs are dealt with by irregular rules (altogether this is not the position they, particularly Pinker in later work, eventually adopt). For them, the competition is *among* the irregular rules the learner postulates: for example, rules R_{S_1} and R_{S_4} (the target) in Figure 4.2.2 may compete to apply to the verb V_1. In the present model, the competition is between an irregular and *the default rule.* Under Pinker and Prince's suggestion, when a target irregular rule loses out, an irregular rule will apply, which results in

Now it should be obvious that we have departed from the Blocking Principle assumed in the WR model (Pinker 1995), a traditional linguistic idea commonly known as the *Elsewhere Condition* (Kiparsky 1973) or the *Subset Principle* (Halle 1997). The Blocking Principle states that when two rules or lexical items are available to realize a certain set of morpho-phonological features, the more specific one wins out. For example, *sang* is used to realized the past tense of *sing*, instead of *singed*, because the former is more specific than the latter, formed by the default -d rule. Call this version of Blocking Principle the *Absolute Blocking Principle* (ABP). In the present model, we suggest a stochastic version of the Blocking Principle (SBP): a more specific rule is indeed *attempted* first before the default rule, but only applies with a probability (its weight). Thus, a more specific rule can be skipped over in favor of a more general rule. The blocking effect of *sang* over *singed* in adult grammar indicates that the weight of the corresponding rule is 1 or very close to 1, *as a result of learning*. In section 4.2.4, we shall return to the Blocking Principle and give empirical arguments for our stochastic version, SBP.

An irregular rule $R_{\mathcal{S}}$, defined over the verb class $\mathcal{S}$, applies with probability $P_{\mathcal{S}}$, once a member of $\mathcal{S}$ is encountered. Thus, it competes with the default -d rule, which could apply to an irregular verb, and in fact does, when $R_{\mathcal{S}}$ does not apply. The acquisition of irregular verb past tense proceeds as in the following algorithm:

Since regular verbs are almost never irregularized, that is, the default -d rule is almost always employed, let's focus our attention on the case where the verb the learner encounters is an irregular one. When presented with a verb in past tense (X_{past}), the learner first reconstructs the root x. As illustrated in Figure 2, the learner then proceeds to analyze the derivation from x to X_{past} in a two step process:

(50) a. associate x to the corresponding class $\mathcal{S}$ and hence the rule $R_{\mathcal{S}}$ defined over this class

b. apply to $R_{\mathcal{S}}$ to x

During learning, neither of the two steps is entirely error-free. First, the learner may not reliably associate x to $\mathcal{S}$, in which case x would be treated as a regular verb (recall that it is

the very rare mis-irregularization errors: the far more abundant overregularization errors, the main fact to explain in our problem, remains unaccounted for.

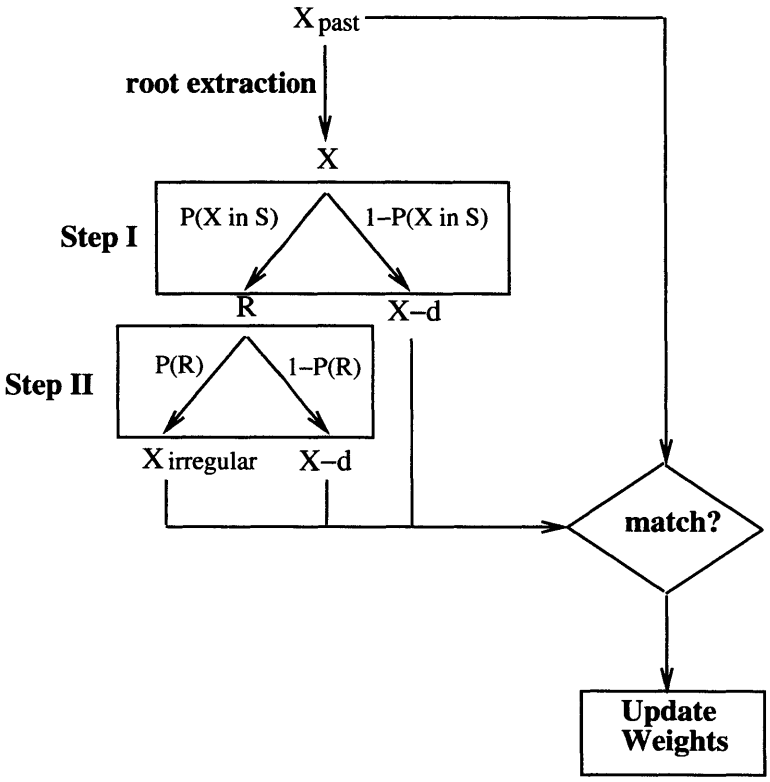

Figure 4-2: Learning irregular verbs by rule competition.

virtually impossible for an irregular verb to misclassified). That is, in (50a), the probability measure $P(x \in \mathcal{S})$ denotes the likelihood that the learner associates x with $\mathcal{S}$. Second, even if x's class membership $\mathcal{S}$ is correctly established, the corresponding rule $R_{\mathcal{S}}$ does not necessarily apply: rather, in (50b), $R_{\mathcal{S}}$ applies with the probability $P_{\mathcal{S}}$, its weight. Only when both decisions are made correctly, the correct past tense will be produced – a match with the input X_{past}. When either of the two decisions fails, the overregularized form will be produced, resulting in a mismatch with X_{past}.

Thus, for each verb, learning involves updating the two probabilities $P(x \in \mathcal{S})$ and $P_{\mathcal{S}}$. Learning is successful when $\forall x, P(x \in \mathcal{S})P_{\mathcal{S}} = 1$: the learner can reliably associate every irregular verb with its corresponding class, and the learner knows with certainty that every irregular rule applies over the default -d rule. As remarked in Chapter 2, many models for updating probabilities (weights) are in principle applicable. For our purpose, let's assume a learner that increases the probabilities of the decisions when they lead to a match between the input form and the analyzed form.

Under the null hypothesis, we assume that the grammar system the child uses for production is the same one that it uses for comprehension/learning, the two-step procedures in the algorithm (50). As a result, overregularization of an irregular verb x occurs when either $P(x \in \mathcal{S}) < 1$ or $P_{\mathcal{S}} < 1$.

More importantly, the RC model makes direct and quantitative predictions about the performance of both irregular verbs and irregular verb classes. Write $\text{CUR}(x) = P(x \in \mathcal{S})P_{\mathcal{S}}$ to denote the *Correct Usage Rate* (CUR) of an irregular verb x. While $P(x \in \mathcal{S})$ may increase when the past tense of x is encountered, $P_{\mathcal{S}}$ may increases *whenever any member* of $\mathcal{S}$ is encountered. These two probabilities, and hence the correct usage of an irregular verb x, are positively correlated with f_x and $f_{\mathcal{S}}$. Hence, if we hold f_x or $f_{\mathcal{S}}$ constant, the RC model makes two directions about the performance of irregular verbs:

(51) a. for two verbs x_1 and x_2 *within* a verb class, $\text{CUR}(x_1) > \text{CUR}(x_2)$ if $f_{x1} > f_{x2}$.

b. for two verbs x_1 and x_2 such that $x_1 \in \mathcal{S}_1$, $x_2 \in \mathcal{S}_2$, and $f_{x1} = f_{x2}$, $\text{CUR}(x_1) > \text{CUR}(x_2)$ if $f_{\mathcal{S}1} > f_{\mathcal{S}2}$.

In section 4.3, we will systematically evaluate these predictions withe children's production data, and demonstrate, in particular, evidence that irregular verbs are indeed organized into

classes.

4.2.4 The Absolute and Stochastic Blocking Principles

We now give justifications for the Stochastic Blocking Principle (SBP), fundamental to the RC model.

Recall that in the WR model, the blocking effect of *sang* over *singed* is given by the ABP: *sang* is used because it is a more specific realization of *sing*+past. The ABP is central to the WR model: when it is presupposed, the rote memorization of irregular verbs is virtually forced. The fact is, children do overregularize, which should be impossible under the ABP. The WR model accounts for this fact by claiming that that irregular verbs are individually memorized. Overregularization errors are explained by appealing to a principle of associative memory: more exposure leads to better memory. The memory imprints of irregular verbs in a child's mind are not as strong as those in an adult's mind, for the simple reason that the child has not seen irregular verbs as many times as adults. Children overregularize because their memory retrieval has yet become reliable.

Pinker (1995, p. 112) justifies the ABP by arguing that the ABP is part of the innate endowment of linguistic knowledge, for it cannot be deduced from its effect. His reasoning is as follows. First, Pinker claims that to learn the ABP, the child must somehow know that forms like *singed* are ungrammatical. Second, it cannot be concluded that *singed* is ungrammatical from its absence in adult speech – absence of evidence does not entail evidence for absence. Finally, Pinker claims that to know *singed* is ungrammatical "is to use it and to be corrected, or to get some other negative feedback signals from adults like disapproval, a puzzled look, or a non sequitur response." Since it is well established (Brown and Hanlon 1973, Marcus 1993; *inter alia*) that children do not have effective negative evidence, it is concluded that the ABP cannot be learned.

It is not the logic of this argument that we are not challenging; rather, it is the premise that the blocking effect of a more specific form over a more general form is absolute. We show that the effect of the blocking in adult language, the motivation for the Blocking Principle in the first place, can be duplicated as a result of learning, without negative evidence, under our stochastic version of the Blocking Principle.

Suppose that, initially, the irregular rule R=[-ø & ablaut] (for *sing-sang*) and the default -d rule are undifferentiated. Upon presentations of *sang* in the input, both rules have a positive probability of being selected to realize *sing*+past. However, only when R is selected can a match result, which in turn increases its weight (probability), P_R. In the end, P_R becomes 1, so that *singed* will never be produced. The end product of such a competition process is a rule system that *appears* to obey the ABP but does not presuppose it: the preference for the specific rule gradually increases, as a result of learning from experience. In the adult system, the default rule simply does not get a chance to apply, for the more specific irregular rule applies first, and with probability 1.

If the effect of the ABP can be duplicated by gradual rule competition and learning, its theoretical status needs to be reconsidered. Our second objection to the ABP is an empirical one. There is at least one good reason to reject the ABP: the presence of *doublets*. For example, *learn*+past can be realized as either *learned* or *learnt*, *dive*+past can be realized as either *dived* or *dove*. For those cases, the ABP cannot be literally true, for otherwise *learned* and *dived* should never be possible, blocked by the more specific *learnt* and *dove*. However, the doublet phenomenon straightforwardly falls out of the SBP with a minor change to the learning algorithm: we suppose that the learner *punishes* P_x when an expected irregular verb x turns out to have regular forms. The term "expected" is important here, implying that the learner has indeed seen irregular forms of x before, but is now being confronted with conflicting evidence. Presumably, speakers that allow both *learned* and *learnt* encounter and use both forms.[7] As a result of competition, the membership probability of *learn* in the corresponding irregular verb class will settle in the interval [0, 1], making alternating forms possible.

4.3 Words vs. Rules in Overregularization

In this section, we examine children's overregularization patterns in detail. We show that the acquisition of irregular verbs shows strong class-based patterns, as predicted by the RC

[7]This includes literary geniuses no less than Lewis Carroll. In *Alice's Adventures in Wonderland*, *learnt* and *learned* appeared exactly once each:

'Yes,' said Alice, 'we learned French and music.'

'Well, I can't show it you myself,' the Mock Turtle said: 'I'm too stiff. And the Gryphon never learnt it.'

model and the rule-based approach to past tense in generative morphophonology.

4.3.1 The Mechanics of the WR Model

In order to contrast the RC model with the WR model, we must be explicit about how the WR model works. For example, for *any* pair of irregular verbs, the RC model makes a concrete prediction about their performance in children (51), based on their input frequencies and the collective frequencies of their respective classes, both of which are directly estimatable from corpora. It is not clear whether predictions can be made with this level of clarity under the WR model. Since irregular verbs are learned by associative pairing in the WR model, it is crucial to expect a *precise* statement of how such associative pairing is established so that the WR model can also take two irregular verbs and predict which one receives a higher CUR in child language. However, the closest to a precise statement in the WR literature is still quite vague (Pinker & Prince 1994: p334):

> It is not clear exactly what kind of associative memory fosters just the kinds of analogies that speakers are fond of. Possibly a network of word-word associations might give rise to the right generalization structure if the design of the lexical representation is informed by modern linguistic theory and its implementation is informed by models of superpositional memory. Here we can only present a rough sketch.
>
> Words might be represented in a hierarchical hardware representation that separates stems and affixes, and furthermore distinguishes foot- and syllable-internal structure, finally representing segmental and featural composition at the lowest level of units. Furthermore each of the possible contents of each representation would be implemented once as a single hardware "type"; particular words would be representation in separate "token" units with pointers to the types it contains. Links between stems and pasts would be set up during learning between their representations at two levels: between the token representations of each pair member, and their type representations at the level of representation that is ordinarily accessed by morphology: syllables, onsets, rhymes, feet (specifically, the structures manipulated in reduplicative and templatic systems, as shown in

> the ongoing work of McCarthy and Prince and others). Ordinary correct retrieval results from successful traversal of token-token links; this would exhaust the process for pairs like *go-went* but would be reinforced by type-type links for members of consistent and high-frequencies families like *sing-sang*. On occasions where token-token links are noisy or inaccessible an retrieval fails, the type-type links would yield an output that has some probability of being correct, and some probability of being an analogical extension (e.g., *brang*). Because the representation of input and output are each highly structured, such extensions would nonetheless be precise and follow constrained patterns, e.g., preserving portions of the stem such as onsets while substituting the appropriate rhymes, and avoiding the chimeras and fuzzy approximations that we do not see among real irregulars but that pure feature-to-feature networks are prone to making.

It is difficult to evaluate the WR model with statements like above. The token level association is clear: the strength of brute force linking between a stem and its past, hence the retrieval rate of the corresponding verb, can be measured by estimating the frequency of the verb's occurrences in past tense. However, it is not clear how the type-level linkings between phonological structures (syllables, onsets, etc.) are established. But far worse is the vagueness of how the two levels interact. For example, while the token-level frequency effect is the dominant factor,[8] it is not clear when the type-level analogy becomes the operative force. Imprecise formulations like the above amenable to analytical results such as (51). It is not even clear whether the WR model supplies enough technical details to carry out a computer simulation on realistic input data.

However, I believe that the evidence presented here is strong enough to preempt *any* classless model, associative mapping or otherwise, from being correct. The acquisition data clearly point to an organization of irregular verbs by rules and classes.

[8] In fact, all the ten pieces of evidence Pinker (1995) offers in support of the WR model, as we shall review in section (4.5), are frequency based, although section (4.3) has shown that frequency affects performance in a fairly subtle way.

4.3.2 The Data

The measure of children's knowledge of irregular verbs is the Correct Usage Rate (CUR). The CUR of an irregular verb x is defined as follows:

$$\text{CUR}(x) = \frac{\text{total number of correct past tense of } x}{\text{total number of past tense of } x}$$

All our data on child performance, unless otherwise noted, come from the monograph *Overregularization in language acquisition* (Marcus et al. 1992), where four American children (Adam 2;3-5;2, Eve 1;6-2;3, Sarah 2;3-5;1, and Abe 2;5-5;0) were studied, using the longitudinal recordings transcribed in the CHILDES corpus (MacWhinney and Snow 1990).[9] Marcus et al. manually analyzed the transcripts, and hence eliminated the unavoidable unambiguity that results from computerized pattern search.[10] The input frequencies of irregular verbs are determined by the present author, based on more than 110,000 adult sentences to which Adam, Eve, Sarah, and Abe were exposed during the recording sessions.

The CURs of all irregular verbs, averaged over all recording sessions, are computed from Marcus et al. (1992, Tables A1-A4) and given in (52):

(52) a. Adam: 2446/2491 = 98.2%

b. Eve: 285/309 = 92.2%

c. Sarah: 1717/1780 = 96.5%

d. Abe: 1786/2350 = 76%

The average CUR for the four children is 89.9%. It is clear that there is quite a bit of individual variation among the children. While Adam, Eve, and Sarah used irregular verbs almost perfectly, Abe's performance is markedly worse. Of particular interest is the verb class [-o & Rime → U], which includes verbs such as *know*, *grow*, *blow*, *fly*, and *throw*. This class posed significant difficulty for all four children. The CURs are 7/16=44% (Adam), 0/1=0% (Eve), 12/22=55% (Sarah), and 28/71=39% (Abe). For Adam, Eve, and Sarah,

[9]Other children Marcus et al. studied are not included here, because of the relatively small size of their recordings and the lack of longitudinal data.

[10]For example, the past tense of no-change irregular verbs can only be accurately identified from the conversation context.

this is the *only* seriously problematic class: all other verbs and classes have close to perfect usage rates. We will explain why this is the case in section 4.3.4.

The WR model learns and organizes irregular verbs on the principle of frequency sensitive associative memory: the more you hear, the better your remember, and the better you retrieve. Hence, CUR(x) for the WR model is directly correlated with the frequency of x in past tense form, f_x. In the RC model, the performance of an irregular verb x is determined by two factors: the probability that x is associated with its class $\mathcal{S}$, and the probability $f_{\mathcal{S}}$ of the rule $R_{\mathcal{S}}$ applying over the default -d rule. Hence, CUR(x) in the RC model is correlated with $f_x \Sigma_{m \in \mathcal{S}} f(m)$.

In what follows, we will examine the two predictions in (51). In particular, we demonstrate the group based patterns in children's production of irregular verbs, which can be directly attributed to rules defined over classes in the RC model, but receives no explanation in the WR model.

4.3.3 Frequency Hierarchy in Verb Classes

The first prediction made by the RC model is straightforward:

(53) for two verbs x_1 and x_2 *within* a verb class, $(CUR)(x_1) > \text{CUR}(x_2)$ if $f_{x1} > f_{x2}$.

To test this prediction, we have listed some verbs grouped by classes in (54), along with their input frequencies estimated from the adult speech.[11] In order to make intraclass comparison, only non-trivial classes are included. Also, to minimize sampling effect, only verbs with at least 20 total occurrences are included in our study (Appendix B gives a complete list of irregular verbs with their frequencies):

(54)

Verbs grouped by class	*Input frequency*
a. [-t & Vowel Shortening]	
lose (80/82=97.6%)	lost (63)
leave (37/39=94.9%)	left (53)

[11] Past tense forms that can unambiguously determined (e.g., *drew*, *took*) were counted by an automated computer search. Ambiguities that arise between past tense and present tense (e.g., *hit*), past participles (e.g., *brought*, *lost*), nouns (e.g., *shot*), and adjectives (e.g., *left*) were eliminated by manually combing through the sentences in which they occurred. Since we are comparing the relative CURs for verbs within a single class, no effort was made to distinguish past tense *put* and *got* from their participle forms, as it is clear that their frequencies thoroughly dominate other members in their respective classes.

b. [-t & Rime → a]

catch (132/142=93.0%)	caught (36)
think (119/137=86.9%)	thought (363)
bring (30/36=83.3%)	brought (77)
buy (38/46=82.6%)	bought (70)

c. [-ø & No Change]

put (239/251=95.2%)	put (2,248)
hit (79/87=90.8%)	hit (66)
hurt (58/67=86.6%)	hurt (25)
cut (32/45=71.1%)	cut (21)

d. [-ø & Vowel Shortening]

shoot (45/48=93.8%)	shot (14)
bite (33/37=89.2%)	bit (13)

e. [-ø & Backing ablaut]

get (1269/1323=95.9%)	got (1,511)
take (118/131=90.1%)	took (154)
write (20/27=74.1%)	wrote (28)
win (20/36=55.6%)	win (36)

f. [-ø & Rime → u]

know (17/23=73.9%)	knew (49)
throw (11/34=32.4%)	threw (28)

Clearly, (54) confirms the prediction in (53): with a single class, the more frequently a verb is heard, the better its CUR.[12] The "exception" in the class (54b), where *think*, a more frequent verb than *catch*, is used at a lower CUR, is only apparent. It is an averaging effect, as (55) makes clear:

(55)

	Children	*Verb*	*% Correct*
a.	Adam, Eve, & Sarah	think	100% (44/44)
		catch	96.5% (110/114)
b.	Abe	think	80.6% (75/93)
		catch	78.6% (22/28)

[12]The strong frequency-CUR correlation in the class [-ø & Backing ablaut] might not be taken at face value. The sound-changing patterns in this class are not homogeneous as other classes, but are nevertheless conventionally labeled altogether as "Backing ablaut". See also footnote (3). The reader is referred to Appendix B for a finer-grained classification of verbs and their CURS.

The low averaged CUR of *think* in (54b) is due to a disproportionately large instances of use from Abe. Once individual variations are factored out as in (55), it is clear that *think* is used correctly at a higher frequency than *catch*, as predicted.[13]

(54) reveals a very important pattern: when and only when verbs are grouped into classes, their performance is, without exceptions, ordered by their input frequencies, which unequivocally points to the conclusion that irregular verbs are organized in (rule-defined) classes. This generalization cannot be captured in theories that do not have verb classes such as the WR model. For example, the frequency-overregularization correlation is also considered by Marcus et al. (1992, p118), who found that for the 19 children tested, the correlation efficient is -0.37 — significant but not without exceptions. In sum, what the WR model shows is that frequency plays an important role in the performance of irregular verbs; what it does not show is the precise manner in which frequency figures into performance.

In order to capture the class-based frequency hierarchy reported in (54), the WR model must duplicate the class-defining effect of rules with "analogy", the type-level association based on phonological similarities of verbs (in a class). Again, it is unfortunate not to have a concrete proposal for analogy in the WR literature. But analogy works only when the sound similarities among verbs under identical rules are strong enough *and* the sound similarities among verbs under different rules are weak enough. A cursory look at the irregular verbs in Appendix A shows this is highly unlikely. For example, verbs in the [-ø & No Change] class, such as *hit, slit, split, quit, spit*, and *bid* are very similar to those in the [=ø & Lowering ablaut] class, such as *sit* and *spit*. Phonological similarity does not give a one-to-one mapping from verbs to classes, and that's why the traditional view in phonology (Chomsky and Halle 1968) treats verb and class association by fiat.

The frequency-performance correlation breaks down when verbs from *different* classes are considered. To see this, we turn to the second prediction made by the RC model, which reveals more empirical problems for the WR model.

[13]We have nevertheless elected to average the performance data, because otherwise the sample size would be too small for each individual child. Furthermore, most errors come from Abe in any case, while the other three children had near perfect use of of all irregular verbs throughout (except for the *ow→ew* class, to which we return momentarily).

4.3.4 The Free-rider effect

Recall that the RC model predicts:

(56) for two verbs x_1 and x_2 such that $x_1 \in \mathcal{S}_1$, $x_2 \in \mathcal{S}_2$ and $f_{x1} = f_{x2}$,
CUR(x_1) > CUR(x_2) if $f_{\mathcal{S}1} > f_{\mathcal{S}2}$.

(56) means that the CUR of an irregular verb x could be quite high even if it is relatively infrequent, as long as other members of its class $\mathcal{S}$ *are* frequently encountered. This "free ride" is made possible by the rule which all members of a class share.

Since most high-frequency verbs are used correctly, we direct our attention to some verbs in (54) that have the lowest input frequencies: *hurt* (25), *cut* (21), *bite* (13) , and *shoot* (14). (We postpone the discussion of *bite* and *shoot* to section 4.3.5 for reasons that will become immediately clear.) We have also included *blew, grew, flew,* and *drew,* which appeared 5, 7, 14, and 22 times respectively, and belong to the [-ø & Rime → u] class that is considerably problematic for all four children.

Consider the six irregular verbs in (57):

(57) *Verbs with very low frequencies (≤ 25 occurrences)*

	Verb Class	*Verbs*	*% Correct*
a.	[-ø & No Change]	hurt, cut	80.4% (90/112)
b.	[-ø & Rime → u]	draw, blow, grow, fly	35.2% (19/54)

Despite the comparable (and low) input frequencies, the verbs in (57a) and (57b) show a sharp contrast in CUR. This is mysterious under the WR model. Since token frequency clearly defies this asymmetry, the WR model must rely on analogy to account for the facts in (57). Again, there is no specific proposal or result to evaluate.

Furthermore, consider the asymmetry between *hurt* and *cut* in (57a) with *know* and *throw* in (54f), the latter of which have higher input frequencies than the former:

(58)

	Verb Class	*Verb (Frequency)*	*% Correct*
a.	[-ø & No Change]	hurt (25), cut (21)	80.4% (90/112)
b.	[-ø & Rime → u]	know (58), throw (31)	49.1% (28/57)

Here the verbs in (58a) are used better than those in (58b), despite of their lower input frequencies. It is not clear how this pattern can be explained by the WR model.

The asymmetries observed in (57) and (58) straightforwardly falls out of the RC model for a simple reason: the *rule* for (57a) and (58a) has much higher weights than those in (57b) and (58b) as a result of learning. The first rule applies the verb *hurt* and *cut*, which do not change in past tense forms. The rule for this class, namely, [-ø & No Change], is amply represented in the input, including *hit, let, set, cut, put*, etc, which have *very* high usage frequencies, totaling over 3,000 occurrences. Every occurrence of such verbs increases the weight of the class rule. Hence, *hurt* and *cut* get a free ride, and have a high CUR despite a low absolute frequency. In contrast, verbs in (57b) belong to the [-ø & Rime → u] class (*blow, grow, know, throw, draw*, and *fly*), which totals only 125 occurrences in our sample of adult input. Hence, the weight of the rule [-ø & Rime → u] must be considerably lower than that of [-ø & No Change]: the CUR asymmetry in (57) is thus accounted.

A closer look at Abe's performance, which is markedly poor across all verb classes, reveals an even more troubling pattern for the WR model. Consider the verbs and their CUR's in (59):

(59) *High frequency verbs with low performance* (Abe)

Class	*Verbs*	*CUR*	*Input frequency*
suppletion	go	.646 (117/184)	557
[-ø & umlaut (ʌ → ey)]	come	.263 (20/76)	272

Verbs in (59) have far higher frequencies than those in (57a), however Abe's performance on them is significantly worse: for the low frequency verbs in (57a), Abe has an average CUR of .659 (29/44, Marcus et al. 1992: Table A8).

This peculiarity in Abe's performance is directly explained by the RC model. Despite their relatively high frequencies, *go-went* and *come-came* nevertheless "act alone", not receiving much help from other members of their respective classes. The suppletion case of *go-went* is obvious. *Come-came* belongs to the heterogeneous class [-ø & umlaut], which in fact consists of three subclasses with distinct sound changes: *fall* and *befall*, *hold* and *behold*, and *come* and *become*. Hence, *come* only receives help from *become*, which isn't much: 2

occurrences in all adult input.[14]

4.3.5 The Effect of Phonological Regularity: Vowel Shortening

Consider the following two low frequency verbs: *shoot* and *bite*, whose past tense forms appeared only 14 and 13 times respectively, in more than 110,000 adult sentences. Despite their rarity in the input, they are used virtually perfectly: 91.8% (78/85) — again in sharp contrast with the performance (40.5%) on the verbs in the [-ø & Rime → u] class (57b).

Past tense formation for both *shoot* and *bite* fall under the rule [-ø & Vowel Shortening]. As remarked in section (4.2.2) and in (48), Vowel Shortening is a pervasive fact of the English language. Furthermore, Myers (1987) shows that Vowel Shortening is essentially "free": vowels in closed syllables are automatically shortened under suffixation, resulting from the interaction between universal phonological constraints and language-specific syllabification properties. Given the evidence that (English) children have good grasp of the syllabic structure of their language (Smith 1973, Macken 1980), learning irregular verbs with Vowel Shortening is considerably simplified, and in fact, reduced to learning which suffix (-t, -ø, or -d) is attached.

In (60), we see that all three classes of Vowel Shortening verbs have very high CUR's:

(60) *Vowel Shortening under suffixation*

	Suffix	*Verb*	*% Correct*	*Input Frequency*
a.	[-t]	lose-lost	98% (80/82)	63
		leave-left	95% (37/39)	53
b.	[-d]	say-said	99% (522/525)	544
c.	[-ø]	shoot-shot	94% (45/48)	14
		bite-bit	90% (33/37)	13

All verbs in (60) are used very well, almost irrespective of their individual frequencies, from very frequent ones (*say-said*) to very rare ones (*shoot-shot*, *bite-bit*). These frequency-defying patterns, along with the asymmetries noted in (57), (57b), and (59), vindicate the reality

[14] Abe's performance on the other two umlaut subclasses are not much better: *fall-fell* is used correctly 72 times out of 129, upon 279 occurrences in the input, and *hold-held* is used correctly 0 of 4 times, upon 11 occurrences in the input, although the sample size in the latter case is too small to be truly informative.

of class-defining morphophonological rules in the RC model. To account for these facts, the WR model must rely on phonological similarities among verbs, which, as remarked earlier and will be further discussed in section 4.4, do not appear to be systematic enough to yield class-based effects.

To sum up this section, we have seen strong evidence for class/rule-based patterns in the learning of irregular verbs. While the correct usage of an irregular verb is related to its input frequency, input frequency is not the only factor. The effect of class-defining rules is also a crucial factor, manifested in free-rider effect (the summed frequencies of *all* the verbs in its *class*; section 4.3.4) and the phonotactics prevalent in the language (section 4.3.5). These facts are readily explained by the RC model, in which morphophonology rules are explicitly employed in the organization and acquisition of irregular verbs.

4.4 Analogy, Regularity, and Rules

4.4.1 Learning by Analogy or Learning by Rules

From the discussion of rule-based performance patterns reviewed in the previous section, we have identified a major problem with the WR model. The regularity among verbs in a class, expressed in a shared phonological rule in the RC model, is not statable in the WR model.

Perhaps the notion of analogy, built on phonological similarity, may duplicate the effect of rules without actually explicitly assuming them. This is the only way to account for the frequency-defying patterns documented in section (4.3). Consider Pinker's discussion on analogy:

> Analogy plays a clear role in language. Children, and adults, occasionally analogize the pattern in one regular verb to a new irregular verb (*write-wrote* → *bite-bote*). They also find it easier to memorize irregular verbs when they are similar to other irregular verbs. The analogizing is a hallmark of connectionist or parallel distributed processing associators; it suggests that human memory might be like a pattern associator. (Pinker 1995: 129)

Could learning by analogy work so that the WR model can be salvaged? Of course one cannot answer this question unless a concrete proposal is given. The question of how

words are analogous to each other, and how analogy is actually used to facilitate learning is usually left vague in the literature, under the rubric of the Wittgensteinian "family resemblance" (e.g., Bybee and Moder 1983, Pinker 1999). Here a methodological point is in order. While there is evidence that some human concepts cluster around fuzzy "family resemblance" categories (Rosch 1978), rather than well-defined classical categories, there is no reason to suppose that words in the linguistic lexicon are necessarily (and exclusively) organized in a similar way. Furthermore, the goal of modern cognitive science is to understand and model mental functions in precise terms. If one is to be content with vague ideas of analogy or association, such as the passage from Pinker and Prince (1994) quoted in section 4.3.1, the systematic regularities among irregular verbs noted in section 4.3, which only surface under close scrutiny of the empirical data guided by a concrete theoretical model, will be swept under the rug unnoticed.

More empirically, there is good reason to suspect that for the acquisition of irregular verbs, learning by analogy cannot be correct in principle. For example, consider the free-rider effect discussed in section (4.3.4), where morphophonological rules enable low frequency verbs to be used with high accuracy: high frequency verbs help strengthen the weight of the rule that all verbs in the corresponding class, high frequency or low, share. In order for the WR model to capture the free-rider effect with analogy, the "family resemblance" between high frequency and low frequency verbs must be very strong. This leads one to expect that the learner will equally strongly "analogize" past tense formation to verbs that *do not* belong to the class but nevertheless *do* bear superficial "family resemblance" with the class members. For example, *bring* may be analogized to *sing* and *ring* to yield *brang*. However, this prediction is empirically false: mis-analogy errors are exceeding rare – about 0.2 percent in all uses of irregular and regular verbs (Xu and Pinker 1995).

Once we move beyond the impoverished morphology of English and on to other languages, it becomes immediately obvious that the WR model cannot be correct. To take an example from Marcus et al. (1995), noun plurals in German employ five suffixes: *Kind-er* (children), *Wind-e* (winds), *Ochs-en* (oxen), Daumen-ø (thumbs; using an empty suffix like the English plural moose-ø and past tense hit-ø), and *Auto-s* (cars). Marcus et al. convincingly argue that despite its low frequency, the -s is the default plural suffix (like the English

-ed for past tense). However, it is hard to imagine that German speakers memorize all four classes of irregular plurals, the majority of nouns in the language, on a word by word basis, as if each were entirely different from the others — this would also be a massive waste of memory, which does not manifest itself very dramatically in English, which has a very small irregular vocabulary. In addition, it is the partial similarity among English irregular verbs that misled Pinker to look for family resemblance:[15] a quick look at German shows that the four irregular classes of noun plurals do not show any systematic similarity whatsoever. Hence, no analogy comes to rescue. It seems that German learners must sort each irregular noun into its proper class, as suggested by the traditional rule-based view. The WR model gets worse when we turn to many African, Pacific, American, and even Indo-European languages with agglutinative morphologies. These languages typically have very long "words" built out of prefixes and suffixes, each of which expresses an individual meaning and all of which are glued together by the grammar and sound system of the language: "he will kiss her" may literally be a concatenation of "kiss", followed by suffixes for future-tense, third person, singular, male, nominative, third person, singular, female, and accusative. Here, too, are default forms: not only past tense for verbs, but also present tense and future, and not only plurals for nouns, but also gender and number. It is inconceivable that these "words", which realize millions or billions of morphological feature combinations, are all individually memorized.

The flaws in the WR model also become clear when other aspects of language acquisition are taken into account. Consider acquisition of agreement morphology, which we briefly reviewed in section 3.2.2. Crosslinguistically, children's agreement morphology is in general near perfect. In addition, we noted in footnote 10 there that their morphological errors are overwhelmingly those of omission (the use of default agreement) rather than substitution (wrong agreement). Now this pattern bears a striking resemblance to children's performance on English irregular verbs: near perfect agreement morphology patterns with near perfect irregular verb morphology, and omission rather than substitution patterns with the rarity of weird past tense errors. This correlation strongly suggests that agreement morphology and past tense morphology are govern by a same mechanism of learning and organization. This

[15]Which seems no more than a historical accident: see section (4.4.2).

is exactly what a general RC model does: the realization of morphological features involves special rules for irregular morphology, and when all fails, the default rule is invoked. To account for this correlation, the WR model must insist that, like English irregular verbs, the systematic agreement morphology in Italian, German, and other languages reviewed in Phillips (1995) is also acquired on the basis of associative learning, which does not seem to be correct.

This is not to say that analogy plays no role in learning. Mis-irregularization errors such as *bring-brang* in children and adult do seem to result from analogy (Prasada & Pinker 1993).[16] However, the role analogy plays in learning must be highly marginal; in fact, as marginal as the rarity of mis-analogy errors, 0.2%, and if almost. This suggests that a very week effect of phonological analogy can be realized in the verb-to-class linking component of the RC model. As for an overall theory of past tense, it is important to note, as Prince and Prince (1988: p127, italics original) remarked, "a theory that can *only* account for errorful or immature performance, without no account of why the errors are errors or how children mature into adults, is of limited value."

4.4.2 Partial Regularity and History

Before moving on, let's consider a major objection that Pinker has raised against the rule-based approach. Since an irregular verb forms past tense by fiat, there is no explanation why verbs like *sting, string, sling, stink, sink, swing*, and *spring* all change *i* to *u* in past participle and all sound so similar (e.g., Pinker 1999, p102, among other places). Pinker's explanation, complete with an oft-cited passage from Wittgenstein, is based again based on family resemblance, the sort of fuzzy associations between stem and past tense in the WR model. Since verbs are represented as bits and pieces of sound segments (Pinker and Prince 1994, Pinker 1999), the common parts they share are reinforced most often and thus become gravitational attraction for word families, with some prototypes close to center such

[16] As pointed out to me by Noam Chomsky and Tom Roeper, an important pattern often ignored is that, by far the most frequent pattern in children's weird past tense errors involve verbs with an *-ing* ending (Xu and Pinker 1995: table 2); see Keyser, Carlson, and Roeper (1977) for an earliest experimental investigation of this issue. Indeed, errors such as *bite-bote* cited by Pinker (1995) and many conceivable errors (e.g., *think-thunk* after *sink-sunk*, *hit-hat* after *sit-sat*) were simply not found. This suggests that irregularization errors are far from a systematic pattern in child language.

as *string-strung* and *sling-slung*, and some on the fringes such as *dig-dug* and *win-won*. But this reasoning seems circular: why are *these* verbs pulled into similarity-based families? As far as one can tell, because they sound similar! Also notice that stem similarity is only partial: the *i-u* family doesn't include *think*, whose past participle is *thought*, or *blink*, which is altogether regular, and both of them seem closer to the family center than *dig* and *win*. Nowhere does the WR model specify us how fuzzy family resemblance actually works to prevent *thunk* and *blunk* from being formed.

The most important reason why Pinker's challenge is not valid is, partial regularity in verb classes is a result of historical contingencies.

In the RC model, verb classes are defined by rules such as (47), repeated below:

(61) *Rule $R_{\mathcal{S}}$ for verb class $\mathcal{S}$:*

$x \xrightarrow{R_{\mathcal{S}}} y$ where $x \in \mathcal{S} = \{x_1, x_2, x_3, \ldots\}$

The members of $\mathcal{S}$ are simply listed, and their correlation is defined by the common rule $R_{\mathcal{S}}$ and/or the common output y. One can imagine another kind of rule that is defined in terms of *input*, where the past tense of the verb is entirely predictable from the stem:

(62) *Rule $R_{\mathcal{S}}$ for verb class $\mathcal{S}$:*

$x \longrightarrow y$ where x has property $\pi_{\mathcal{S}}$.

In present-day English, rules like (62) are full of exceptions. However, their regularities were higher (though perhaps never complete) further back in history. Even the suppletive verbs, which may seem arbitrary synchronically, are nevertheless non-accidental diachronically. For example, why do we have *go-went*, seemingly unrelated, but not *go-smeeb*, also seemingly unrelated? In Middle English, *go* somehow replaced the now obsolete *wend*. However, *go* did retain the past tense form, *went*, which belongs to the more regular class that also includes *bend* and *send*. Hence, the suffixation and readjustment rules, synchronically productive, are evidenced diachronically: no irregular verbs are exception to -t, -ø, and -d suffixation.

How did such (partial) regularities get lost in history? There are two main factors (cf. Pinker 1999: Chapter 3). One is purely frequency-based. If an irregular verb is used very infrequently, the learner will not reliably locate it in the appropriate class to which it belongs. We will return to this in section (4.5.9). The other factor falls out of the interaction between

irregular rules and changes in other parts of the phonological system. A perfect example is given by Pinker (1999: 65-66). The word *wrought*, the archaic past tense of *work* that belongs to the *bring* class, appears quite mysterious to many contemporary speakers. The mystery is explained historically in the following way. First, adding the suffix -t to *work* yielded *workt*. Second, [*k*] was softened to [*gh*], as in *Bach*, to yield *worght*. Third, a vowel and an adjacent *r* (a vowel-like liquid) are often switched during the history of English: *brid-bird*, *hross-horse*, and *worght-wroght*. Fourth, *gh* stopped being pronounced, and, the [aw] became [o] as the result of the Vowel Shortening rule upon suffixation, which, as we have noted earlier, falls out the interaction of syllabification and UG principles. Other words in the *bring-brought* class can be similarly explained. Again, we see that the natural history of irregular verbs is not completely random, but rather stochastic: sampling effects and other unpredictable changes, such as *go* replacing *went*, interact with predictable UG principles and conventions to produce partial similarities observed in irregular verb classes.

4.5 Some Purported Evidence for the WR Model

Pinker (1995) summarizes previous work on the WR model and gives ten arguments to its support; see Pinker (1999) for a popular rendition. Here we review them one by one, and show that they are either factually inaccurate, or methodologically flawed, and in case of either, they are handled equally well by the RC model.

4.5.1 Error Rate

How low is it?

Pinker claims that the rate of past tense errors is quite low: the mean rate across 25 children is 4.2%, the median only 2.5%. He suggests that the low error rate indicates that overregularization is "the exception, not the rule, representing the occasional breakdown of a system that is built to suppress the error", as in the WR model.

First, it is questionable whether the actual error rate is actually *that* low. In (52), we have seen that the error rate averaged over four children is 10.1%. In particular, Abe's error rate is *very* high: about 24% of the irregular verbs were regularized. Also, as it is clear from

Table A8 in Marcus et al. (1992), Abe's poor performance cuts across all verb classes, and thus is not due to a few particularly bad but very frequent verbs/classes. He even made a considerable number of errors (64/177=36%) in *go-goed*, while all other children used *went* perfectly throughout. Second, by averaging over all irregular verbs, the more problematic but less frequent verbs and classes, and the important variations *among* classes (section 4.3) are lost. For example, *all* four children performed very badly on the [-ø & Rime → u] class, an error rate of 48.6% (54/111).

Longitudinal trends

Pinker claims that the rate of overregularization, 2.5%, is stable through the preschool years (2 to 5), and gives Adam's longitudinal overregularization trend, which is indeed quite steady (and low) over time. He concludes that the steady error rate is due to the occasional malfunction of memory retrieval – the exception, not the rule.

There is strong reason to challenge this claim. First, it seems that *Adam* is the exception, rather than the rule. Adam's grasp of irregular verbs is in general perfect, the best among the four children we examined; see (52). Second, as already noted in section 4.5.1, averaging over all irregular verbs is likely to obscure longitudinal patterns, which could be observed only with problematic verbs for children.

Fortunately, we do have Abe, whose irregular verb performance is, across all verb classes, markedly worse than the other three children. To study Abe's longitudinal development, we have grouped every consecutive 15 recordings into a period. There are 210 recordings (from 2;4 to 5;0), so we have 14 periods altogether. We have examined verbs that Abe was particularly bad at: *go, eat, fall, think, came, catch, run*, and the members of the problematic [-ø & Rime → *u*] class: *throw, grow, know, draw, blow*, and *fly*. The results are are summarized in the table below:

With the exception of period 1, in which Abe only had 18 opportunities to overregularize (and thus a likely sampling effect), his error rate is no doubt gradually declining. This shows that children's overregularization at the earliest stage is considerably more systematic than Pinker claims, and cannot be attributed to simple performance errors.

Period	# of Overregularization	*Total # of Use*	*Error Rate*
1	3	18	0.167
2	14	25	0.560
3	31	50	0.620
4	27	37	0.729
5	10	19	0.526
6	28	56	0.500
7	28	54	0.519
8	7	38	0.184
9	18	52	0.346
10	10	40	0.250
11	4	33	0.121
12	4	23	0.174
13	2	43	0.047
14	3	46	0.065

Table 4.1: Abe's longitudinal overregularization for problematic verbs.

4.5.2 The role of input frequency

Pinker notes that the more frequently an irregular verb is heard, the better the memory retrieval for that verb gets, and the lower the overregularization rate is. This claim, while correct for verbs within a class (section (4.3.3)), is in general incorrect. The CUR of an irregular verb is determined by two factors, the correct identification of class membership, and the weight of the irregular rule. In sections (4.3.4) and (4.3.5), we have seen that verbs like *shoot*, *bite*, *hurt*, and *cut*, which appear infrequently in adult speech are used almost perfectly, whereas verbs such as *draw, blow, grow*, and *fly*, which are also comparably infrequent, fare much worse.

4.5.3 The postulation of the -d rule

In the stage which Pinker calls phase 1 (from 2;3 to shortly before 3;0), Adam left many regular verbs unmarked: instead of saying *Yesterday John walked*, the child would say *Yesterday John walk.* Overregularization started in phase 2, as the rate of tensed verbs very rapidly became much higher. Pinker suggests that the two phases are separated by the postulation of the -d rule. Although this appears to be a reasonable interpretation, it is

problematic when individual variations and other aspects of language acquisition are taken into consideration.

First, individual variations. Pinker (1995) only gives the tendency of regular verb marking for Adam, based on Marcus et al. (1992: 109). However, also in Marcus et al. (1992: 109-111), we see that the other three children showed very different patterns. Eve's use of regular verbs was basically in a steady climb from the outset (1;6). Sarah showed quite a bit of fluctuation early on, perhaps due to sampling effect, before gradually settling on an ascent. Abe, whose irregular verbs were marked poorly, nevertheless showed the highest rate of regular verb marking: he started out with about 70% of regular verb marking at 2;5, which went to 100% around 2;10.

Second, the low rate of tense marking in phase 1 has been known in the acquisition literature as the Optional Infinitive (OI) stage, first reported by Weverink (1989) and Pierce (1989/1992). Children in the OI stage produce a large amount of non-finite verbs in matrix sentences as well as finite ones. This observation has been made in the acquisition of Dutch, English, French, German, and other languages. Since the non-finite root verbs are extensively used, the consensus in the field is that, for some reason, non-finite matrix verbs reflect part of the child's grammatical system. For example Rizzi (1992) has found that in French, German, and Dutch, sentences with Wh fronting almost always involve finite verbs.[17] Roeper and Rohrbarher (1994) and Bromberger and Wexler (1995) have found that in English Wh-questions, null subjects almost always correlate with OI verbs (*where going?*), and overt subjects almost always correlate finite verbs (just a handful of examples like *where you going?*). Schütze (1997), among others, found that case errors significantly correlate with tense: default case errors (*me going*) almost always involve OI verbs, where correct (nominative) errors almost always involve finite verbs. These results strongly point to some grammatical mechanism studying the OI phenomenon, not merely a performance effect.

Pinker's proposal in effect reduces the OI stage to the lack of the -d rule early on, but this becomes problematic when the OI phenomenon is carefully considered. For example, in languages with verb raising, e.g. French, verbs in higher clausal positions (TP) *are* generally

[17]See Wexler (1994, 1998) for a summary of OI research.

inflected, and verbs in base positions are generally not inflected. The following examples from child French are taken from Pierce (1992: 65), where the tense and verb-raising correlation is indicated by the negation marker *pas*:

(63) a. [-finite]
pas manger
not eat

b. [+finite]
veux pas lolo
want not water

This asymmetry with respect to verb position is not predicted if the finiteness of verbs is simply a matter of availability and application of the -d rule. Also, there are languages (e.g. French, but not English) for which the root form and the infinitive form of verbs are morphologically and phonologically different. In the OI stage of such languages, it is the *infinitive* forms that are used (63a), not the root forms. Pinker's proposal predicts that root forms are used instead of past tense forms, due to the lack of the -d rule – this is contrary to factual findings.

Consider an alternative analysis for the rapid increase Pinker noted in the use of inflected verbs. No discontinuity is supposed of the -d rule; that is, we assume that the -d rule is available to the learner quite early on. However, during the OI stage, the -d rule, which applies to past tense verbs, simply does not apply to the extensively used non-finite verbs that are allowed by an OI stage competence system. When children leave the OI stage, the -d rule consequently becomes applicable.

Now this alternative proposal makes two immediate predictions. First, a child's exit from the OI stage ought to coincide with the increase of past tense verbs in its production.[18] This is essentially Pinker's observation in Adam. Second, there are languages such as Italian and Spanish that do *not* have the OI stage in acquisition, and verbs are almost always inflected (Guasti, 1992). If the alternative view is correct, that the -d rule is available from early on, we predict that in the acquisition of Italian and Spanish, irregular verbs

[18]This prediction must be checked against the longitudinal data of a *single* child, since there is quite a bit of individual variation among children's OI stage, and there is also quite a bit of individual variation among children's verbal past tense, as we have seen.

ought to be overregularized also from early on. The late postulation of the -d rule in the WR model does not make this prediction. We have not checked this prediction, which might further distinguish the two proposals. Elicitation experiments could also be conducted to test whether children that are supposed to lack the -d rule during phase 1 would overregularize when the root verb must be finite under controlled conditions (e.g. higher position in the clause).

4.5.4 Gradual improvement

Pinker notes that after the -d rule is postulated (but see section (4.5.3) for an alternative view), overregularization does not drive out correct use of irregular verbs, but bare forms instead, which are extensively used during phase 1. He cites Adam's performance for support. Adam's average CUR is .74 during phase 1, and .89 during phrase 2. There appears to be no "real regression, backsliding, or radical reorganization" (1995: 118) in Adam's irregular verb use. This follows if the memory for irregular verbs is getting better.[19]

However, the gradual improvement pattern is also predicted by the RC model, as weights for class membership and irregular rules can only increase. The gradual improvement in the performance results from the increasing amount of exposure to irregular verbs.

4.5.5 Children's judgment

Experiments have been conducted to test children's knowledge of irregular verbs, by presenting children with overregularized verbs and asking them if they sound "silly". Children are found to call overregularized verbs silly at above chance level. This finding is claimed to show that children's grammar does deem overregularization worse, despite their occasional use.

Pinker correctly points out some caveats with such experiments: children's response might be affected by many factors, and is thus not very reliable. In any case, these findings are hardly surprising: even Abe, the child with by far the worse irregular verb use, had an overall error rate of 24% – far better than chance. In fact, such findings are compatible with

[19]The gradual improvement in Adam's performance seems to contradict Pinker's earlier claim that Adam's error rate is stable (section (4.5.1)).

any model, including the present one, under which children produce more correct forms than overregularizations at the time when the experiment was conducted.

4.5.6 Anecdotal evidence

Pinker cites two dialogues (one is given below) between psycholinguists and their children, during which the adults use overregularized verbs to observe the children's reaction. The children are not amused.

> Parent: Where's Mommy?
>
> Child: Mommy goed to the store.
>
> Parent: Mommy goed to the store?
>
> Child: NO! (*annoyed*) Daddy, *I* say it that way, not you.

Pinker (1995: 119) suggests that the children, "at some level in their minds, compute that overregularizations are ungrammatical even if they sometimes use them themselves."

Whether anecdotal evidence should be taken seriously is of course a concern here. But an immediate question comes to mind: the fact that children don't like adults using overregularization may be due to their perceptions of adults, *as adults*, who are expected to say things differently, or, they are simply annoyed at being repeated. In any case, the RC model gives a more direct explanation for observed reactions. Recall that at the presentation of each past verb, the child has probabilistic access to either the special irregular rule (when applicable), or the default -d rule, to generate the expected past tense form from the extracted root. Now if an overregularized form such as *goed* is repeated several times, the chance of a mismatch (i.e. the child generating *went*) is consequently enhanced – the probability of generating *went* at least once in several consecutive tries – much to children's annoyance, it appears.

4.5.7 Adult overregularization

Adult do occasionally overregularize. Pinker claims that the rarity entails that adult overregularization is the result of performance, but not the result of a grammatical system.

However, this is not the only interpretation of adult overregularization: rule-based grammatical system approaches account for the data equally well. Under the RC model, for an irregular verb (e.g. *smite-smote*) that appears very sparsely, the learner may not sure which class it belongs to, i.e., the probability of class membership association is considerably below 1. Overregularization thus results, even if the weight of the irregular rule for its corresponding class is infinitely close to 1.

Pinker also notes that since memory fades when people get old, more overregularization patterns have been observed during experiments with older people.[20] This interesting finding vindicates all theories that handles irregular verbs learning with some form of memorization: in the RC model, it is the class membership that is memorized.

4.5.8 Indecisive verbs

Adults are unsure about the past tense of certain verbs that they hear infrequently. *Dreamed* or *dreamt*? *Dived* or *dove*? *Leapt* or *leaped*? *Strided* or *strode*?[21]

Pinker links input frequency to the success of irregular past tense (memory imprint). Again, this correlation is also expected under the RC model: low frequency verbs give the learner little clue about class membership, and for doublets, the class membership is blurred by the non-trivial frequencies of both forms.

4.5.9 Irregulars over time

Pinker cites Joan Bybee's work that of the 33 irregular verbs during the time of Old English, 15 are still irregular in Modern English, with the other 18 lost to the +ed rule. The surviving ones had a frequency of 515 uses per million (137/million in past tense), and the regularized ones had a frequency of 21 uses per million (5/million in past tense). The more frequently used irregulars are retained.

The RC model readily accounts for observation. Suppose that for generation n, all 33 irregular verbs had irregular past tense forms, but some of them are very infrequently used.

[20] More specifically, patients with Alzheimer's and other largely age-related diseases (Ullman et al. 1993; cf. Ullman et al. 1997).

[21] Some of those forms are doublets, so both forms are heard. As noted in section 4.2.4, they pose a problem for the Absolute Blocking Principle, on which the WR model is built.

As a result, generation $n+1$ will be unsure about the class membership of the infrequent irregulars, for reasons discussed in section 4.5.8, and will regularize them sometimes. Consequently, generation $n+2$ will be even less sure and will produce more regularized forms. Eventually, when the irregular forms will drop into non-existence, such verbs will have lost their irregular past tense forever. Thus, the loss of irregularity is a result of sampling effects and competition learning over time.

4.5.10 Corpus statistics

Based on the statistics from modern English text corpora, Pinker found that the top 10 most frequently used verbs are all irregular verbs, and that 982 of the 1000 least frequently used are regular verbs. He reasons that this pattern is predicted since the survival of irregular verbs against children and adults' overregularization is only ensured by high frequency of use. This is certainly correct, but is also obviously compatible with the RC model, following the discussion in 4.5.8 and 4.5.9.

4.6 Conclusion

In sum, we have proposed a novel Rules and Competition model for the acquisition of past tense in English. A list of probabilistic suffixation and readjustment rules, defined over classes of irregular verbs, compete with the default -d rule for past tense inflection. Hence, the learning of an irregular verb is determined by the probability with which it is associated with its class, *and* the probability with which the class rule applies over the default -d rule. We have also given justifications for, and explored the consequences of, a stochastic and learning-theoretic version of the Blocking Principle.

On the basis of children's overregularization errors, we have developed a critique of the two module Word and Rule model. The differences between the two models lie in how the irregular verbs are organized and used. For the WR model, irregular verbs are stored as associated pairs. This leaves unexplained many systematic regularities in irregular verbs such as the class-based frequency hierarchy, the free-rider effect and the role of language-specific phonotactics, all strongly attested. For the RC model, such regularities are captured

via shared rules, which compete against the default -d rule and gradually grow in strength, as a result of learning from experience.

Appendix A: The Rule System for English Past Tense

This list is loosely based on Halle & Mohanan (1985: Appendix) and Pinker & Prince (1988: Appendix). Extremely rare verbs are not listed.

suppletion

go, be

-t suffixation

- No Change

  burn, learn, dwell, spell, smell, spill, spoil

- Deletion

  bent, send, spend, lent, build

- Vowel Shortening

  lose, deal, feel, kneel, mean, dream, keep, leap, sleep, leave

- Rime $\rightarrow$ *a*

  buy, bring, catch, seek, teach, think

-ø suffixation

- No Change

  hit, slit, split, quit, spit, bid, rid, forbid, spread, wed, let, set, upset, wet, cut, shut, put, burst, cast, cost, thrust, hurt

- Vowel Shortening

  bleed, breed, feed, lead, read, plead, meet

hide, slide, bite, light

shoot

- Lowering ablaut

  sit, spit, drink, begin, ring, shrink, sing, sink, spring, swim

  eat, lie

  choose

- Backing ablaut

I $\rightarrow \wedge$	fling, sling, sting, string, stick, dig, win
ay $\rightarrow$ aw	bind, find, grind, wind
ay $\rightarrow$ ow	rise, arise, write, ride, drive, strive, dive
ey $\rightarrow$ U	take, shake
er $\rightarrow$ or	bear, swear, tear, wear
iy $\rightarrow$ ow	freeze, speak, steal, weave
$\varepsilon \rightarrow a$	get, forget

- umlaut

  fall, befall

  hold, behold

  come, become

- V $\rightarrow$ u

  blow, grow, know, throw, draw, withdraw, fly, slay

-d suffixation

- Vowel Shortening

  flee, say

- Consonant Deletion

  have, make

- ablaut

  sell, tell

- No Change (default)

  regular verbs

Appendix B: Overregularization Errors in Children

Irregular verbs are listed by classes; in the text, only verbs with 25 or more occurrences are listed. The counts are averaged over four children. All raw counts come from Marcus et al. (1992), with one exception noted in footnote.

- [-t & Vowel Shortening]

  lose 80/82, feel 5/18, mean 4/5, keep 2/2, sleep 3/6, leave 37/39

- [-t & Rime $\rightarrow$ *a*]

  buy 38/46, bring 30/36, catch 132/142, teach 8/9, think 119/137

- [-ø & No Change]

  hit 79/87, cut 32/45, shut 4/4, put 239/251, hurt 58/67

- [-ø & Vowel Shortening]

  feed 0/1, read 1/2, hide 4/5, bite 33/37, shoot 45/48

- [-ø & Lowering ablaut]

  sing 3/4, drink 9/15, swim 0/3, sit 5/7, spit 0/3

  eat 117/137

- [-ø & Backing ablaut]

  stick 5/10, dig 2/5, win 20/36

  ride 7/8, drive 6/12

  take 118/131, shake 4/4

  get 1269/1323, forget 142/142

- [-ø & umlaut]

  fall 266/334

  hold 0/5

  come 109/174,

- [-ø & Rime → u]

  blow 5/15, grow 4/12, know 17/23, throw 11/34, draw 2/12, fly 8/15

- [-d & Vowel Shortening]

  say 522/525

Chapter 5

Internal and External Forces in Language Change

> An observed linguistic change can have only one source — a change in the grammar that underlies the observed utterances.
>
> Noam Chomsky and Morris Halle
> *The Sound Patterns of English* (1968, p249)

Language change is observed when one generation of speakers produces linguistic expressions that are different from previous generations, either in form or in distribution.[1] Language change is explained when its causal forces are identified and their interactions are made clear.

At least two components are essential for any causal theory of language change. One component, long recognized by linguists (Halle 1962, Chomsky and Halle 1968, Lightfoot 1979, *inter alia*), is a theory of language acquisition by child learners: ultimately, language changes because learners acquire different grammars from their parents. In addition, as children become parents, their linguistic expressions constitute the acquisition evidence for the next generation. Following Battye & Roberts (1995), this iterative process can be stated in the familiar distinction between E- and I-languages (Chomsky, 1986):

[1] I bear a special debt to Ian Roberts, Tony Kroch, and Ann Taylor for their careful work on the history of French and English, on which all my empirical results in this Chapter rest.

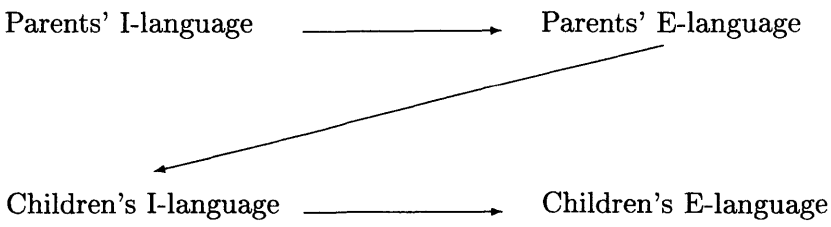

Figure 5-1: The dynamics of language acquisition and language change

The other crucial component in language change became clear through the generative linguistics research of the past half century. The restrictiveness of human language, amply reflected in child language acquisition, coupled with the similarities revealed in comparative studies of the world's languages, have led many linguists to conclude that human languages are delimited in a finite space of possibilities, the P&P framework. A Universal Grammar (UG) is proposed as part of our biological endowment, which consists of discrete rules and constraints that interact in infinite yet non-arbitrary ways. Therefore, language acquisition, and hence language change, are determined by both internal and external factors: the internal knowledge of the Universal Grammar determines the space of languages that learners *can* attain, and the external linguistic experience in the environment determines what language children *do* attain. Their interactions, as depicted in Figure 5, in turn determines the space of language change.

The causal role of language acquisition in language change dictates that a model of language acquisition be at the heart of any explanatory model of language change. When one gives descriptions of a certain historical change, for example, the change of a parameter from one value to another, one must give an account, from a child language perspective, of *how* that change took place. Hence, all the empirical conditions imposed on an acquisition model outlined in Chapter 1, must apply to a language change model with equal force. Of these, two aspects deserve particular attention.

First, the model must make *quantitative* predictions about the direction of language change at time $t + 1$ and beyond, when presented with the composition of linguistic data time t. For example, one would like to make claims that when such and such patterns are found in the historical text, such and such changes are bound to occur.

Second, one must follow the condition of explanatory continuity in studying language change. It is common to find in the literature appeals to socio-political factors to explain language acquisition. However, this approach is not complete unless one develops a formal, quantitative, developmentally compatible, and independently motivated model which details how such factors affect language acquisition, the causal force in language change. It is also common to find notions such as "diachronic reanalysis", which claims that the learner under certain conditions will opt for a radical change in his grammar. Again, these claims can only be substantiated when supporting evidence can be found in synchronic child language development.

This chapter extends the acquisition model to a study of language change that aims to meet the conditions above. It characterizes the dynamical interaction between the internal Universal Grammar and the external linguistic evidence, as mediated by language acquisition. We will again borrow insights from the study of biological evolution, where internal and external forces, namely, genetic endowment and environmental conditions, interact in a similar fashion. Section 5.1 lays out the details of the language change and derives a number of formal properties, including a sufficient and necessary condition under which one grammar replaces another. In sections 5.2 and 5.3, we apply the model to explain the loss of verb second (V2) in Old French and the erosion of V2 in Old English.

5.1 Grammar Competition and Language Change

5.1.1 The role of linguistic evidence

The fundamental question in language change is to identify the causal forces that make generation $n+1$ attain knowledge of language that is different from generation n.

Recall that the language attained by a learner is the product of internal knowledge of Universal Grammar and external linguistic evidence present in the environment, which are mediated by the algorithm of language acquisition. If we assume, as there is no reason otherwise, that the biological endowment of Universal Grammar is held constant from generation to generation, we may conclude that the only source for the discrepancy between two generations of speakers must lie in the linguistic evidence: generation n and $n+1$ are exposed

to sufficiently different linguistic evidence and thus form different knowledge of language as a result.

This conclusion is only warranted under some further justifications. We will argue that language change cannot take place without sufficiently different linguistic evidence across generations. With a generation of speakers viewed a population of individuals, it remains a theoretical possibility that in spite of comparable linguistic evidence, *some* members of generation $n+1$ attain a different grammar from generation n, as a result of imperfect *mis-learning*. However, this position is empirically untenable in three ways. First, language acquisition research shows that children are highly competent and robust learners: it seems improbable that given sufficiently similar experience, children will attain languages that are substantially different (e.g., a major syntactic parameter is set to a wrong value in a significant proportion of the population). Second, historical linguistics shows that language change occurs on the scale of the entire population, not scattered individual members, as Bloomfield (1927, cited in Hockett 1968) comments:

> It may be argued that change in language is due ultimately to the deviations of individuals from the rigid system. But it appears that even here individual variations are ineffective; whole groups of speakers must, for some reason unknown to us, coincide in a deviation, if it is to result in a linguistic change. Change in language does not reflect individual variability, but seems to be a massive, uniform, and gradual alteration, at every moment of which the system is just as rigid as at every other moment.

Third, while one might attempt to invoke the idea of individual mis-learning to explain historical change in *some* languages,[2] it leaves mysterious the relative stability in *other* languages, say, the rigidity of word order in Western Germanic languages.

We therefore reject mis-learning (under sufficiently similar linguistic evidence) as a possible mechanism of language change. A question immediately arises: What makes the linguistic evidence for generation $n+1$ different from that of previous generation? There are many possibilities. For example, migration of foreign speakers might introduce novel

[2]Indeed, this is the approach taken by Niyogi and Berwick (1995), whose model of language change relies on mis-convergence by triggering learners in the sense of Gibson and Wexler (1994).

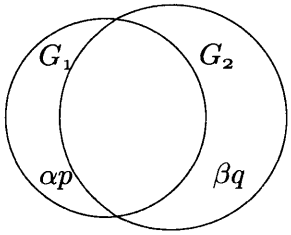

Figure 5-2: Two mutually incompatible grammars constitute a heterogeneous linguistic environment.

expressions that were previously unseen; social and cultural factors might also eschew the distributional patterns of linguistic expressions used in a population. These are interesting and important topics of research, but are not relevant for a formal model of language change.[3] Hence, we are chiefly concerned with the *predictable consequences* of such changes: what happens to language learners when the linguistic evidence is altered, and how does it affect the composition of the linguistic population as a result?

5.1.2 A Variational Model of Language Change

Suppose that, resulting from migration, genuine innovation, and other sociological and historical factors, a linguistic environment is established for a generation of language learners that is substantially different different from the one for the previous generation. The expressions used in such an environment, call it E_{G_1,G_2}, can formally be viewed as a mixture of expressions generated by two independent sources: the two grammars G_1 and G_2. Further, suppose a proportion α of G_1 expressions are incompatible with G_2, and a proportion β of G_2 expressions are incompatible with G_1. Call α (β) the *advantage* of G_1 (G_2). The following figure illustrates:

The variational approach views language acquisition as competition and selection among grammars. Recall that the fitness of individual grammars is defined in terms of their *penalty*

[3]This is much like the population genetic theory of natural selection, which concerns the predictable changes in the population once some new genotypes are introduced: the precise manner in which new genes arise, which could be mutation, migration, etc., is a separate question to which often contains too much contingency to demand a firm answer.

probabilities:

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

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

The penalty probabilities ultimately determine the outcome of language acquisition:

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

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

Suppose that at generation n, the linguistic environment $E_{G_1,G_2} = pG_1 + qG_2$, where $p + q = 1$. That is, in E_{G_1,G_2}, a proportion p of expressions are generated by G_1, and a proportion q of expressions are generated by G_2, and they collectively constitute the linguistic evidence to the learners in generation $n+1$. The penalty probabilities of G_1 and G_2, c_1 and c_2, are thus βq and αp. The results in (65) allow us to compute p' and q', the weights of G_1 and G_2 respectively, that are internalized in the learners of generation $n+1$:

(66) The dynamics of a two grammar system:

$$\begin{aligned} p' &= \frac{\alpha p}{\alpha p + \beta q} \\ q' &= \frac{\beta q}{\alpha p + \beta q} \end{aligned}$$

(66) shows that an individual learner in generation $n+1$ may form a combination of two grammars G_1 and G_2 at a different set of weights than the parental generation n.[4] From (66), we have:

$$\begin{aligned} \frac{p'}{q'} &= \frac{\alpha p/(\alpha p + \beta q)}{\beta q/(\alpha p + \beta q)} \\ &= \frac{\alpha p}{\beta q} \end{aligned}$$

[4] Suppose, in a uniform linguistic environment E_{L_1} a small number (n), out of a total of N learners, *do* mis-converge to a non-target grammar L_2. The effect of the mis-learners on the next generation can be quantified:

$$E_{L_1,L_2} = L_1 \frac{N-n}{N} + L_2 \frac{n}{N}$$

If $n << N$, then the linguistic environment for the next generation is virtually identical to a uniform environment without mis-learners. Thus, the impact of the mis-learners on the next generation is negligible.

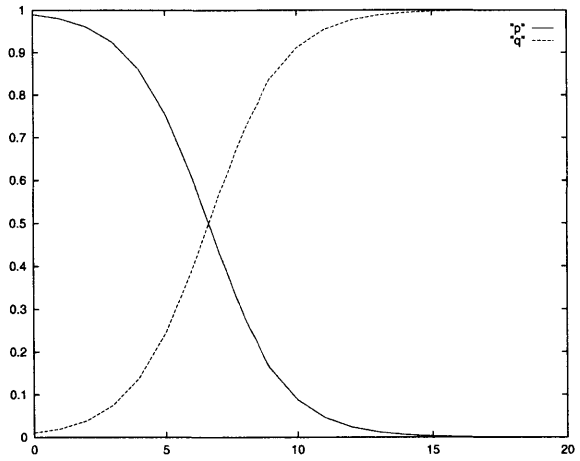

Figure 5-3: One grammar (q) replacing another (p) over time.

In order for G_2 to overtake G_1, the weight of G_2 (q) internalized in speakers must increase in successive generations and eventually drive the weight of G_1 (p) to 0. That is, for each generation, it must be the case that $q' > q$, which is equivalent to $p'/q' < p/q$. Thus, we obtain a sufficient and necessary condition for grammar competition in a linguistic population:

(67) *The Fundamental Theorem of Language Change*

G_2 overtakes G_1 if $\beta > \alpha$: the advantage of G_2 is greater than that of G_1.

Recall that α and β are presumably constants, which characterize the distributional patterns in the use of the respective languages. Note that we may not be able to estimate α and β directly from historical context, which only reflects the penalty probabilities of the competing grammars, i.e. αp and βq. However, (67) says that *if* $q' > q$ (G_2 is on the rise), it must be the case that $\beta > \alpha$, *and*, if $\beta > \alpha$, G_2 will necessarily replace G_1. Hence, we have the following corollary:

(68) Once a grammar is on the rise, it is unstoppable.

Plotting the $q(t)$, the weight of G_2, as a function of time t, we obtain the familiar S-shape curve (see Figure 5.1.2) that has often been observed in language change (Weinreich et al. 1968, Bailey 1973, Kroch 1989, Keenan 1998, among many others), as the "new" linguistic form gradually replacing the "old" form:

The present model shares an important feature with Clark & Roberts' (1993) work, which extends the use of Genetic Algorithms in acquisition (Clark 1992). In both models, the outcome of language acquisition is determined by the compatibilities of grammars with linguistic evidence, in a Darwinian selectionist manner. However, they identify the final state of acquisition with a single grammar (cf. Niyogi and Berwick 1995). Therefore, when the linguistic evidence does not unambiguously identify a single grammar, as a realistic, inherently variable environment, they posit some general constraints on the learner, e.g. the *elegance* condition, which requires the learner to select the simplest among conflicting grammars. Aside from such explanatorily discontinuous assumptions that require independent justification, the position of learner converging to a single grammar cannot be defended in face of the empirical evidence found by Kroch and his colleagues (Kroch 1989, Pintzuk 1991, Santorini 1992, Kroch and Taylor 1997, Kroch, Taylor, and Ringe 1997). They have shown that historical texts during the period of In fact, historical linguists commonly use terms such as "erosion" or "optional application" to indicate the gradual appearance of a grammatical construction. These facts, and more generally, linguistic variability of the sort noted by Weinreich, et al. (1968), can straightforwardly be modeled as co-existence of multiple UG grammars in the approach taken here.

For the purpose of this study, we assume that all speakers in a linguistic community are exposed to identical linguistic experience, and that speaker's linguistic knowledge is stable after the period of language acquisition (i.e. there is no generational overlap). It is possible to incorporate such spatially and temporally varying factors into the dynamics of language change, which may be aided by the well-established models of population genetics and evolutionary ecology. We leave these options for further research.

To summarize the theoretical considerations in this section, we have extended the variational model of language acquisition to a population of learners and presented some analytical results concerning the dynamical system thus construed. We conclude that heterogeneity in the linguistic evidence, however introduced, is a prerequisite for language change. Once the homogeneity is punctured, language learners form internal representations of co-existing grammars. The propagation of such grammars in successive generations of individual learners defines the dynamics of language change. We now apply the variational model of lan-

guage change to the loss of V2 in the history of French and English, drawing comparison and connection to previous analyses.

5.2 The Loss of V2 in French

Old French (OF) had a cluster of properties, including V2 and *pro*-drop, that are lost in Modern French (ModF). The following examples are taken from Clark and Roberts (1993):

(69) *Loss of null subjects*:

a. *Ainsi s'amusaient bien cette nuit. (ModF)
thus (they) had fun that night.

b. Si firent grant joie la nuit. (OF)
thus (they) made great joy the night.

(70) *Loss of V2*:

a. *Puis entendirent-ils un coup de tonnerre. (ModF)
then heard-they a clap of thunder.

b. Lors oïrent ils venir un escoiz de tonoire. (OF)
then heard they come a clap of thunder.

In this section, we will provide an analysis for the loss of V2 under the variational model. All examples and statistics cited in the remainder of this section are taken from Roberts (1993, henceforth R).

Recall that in order for a ModF SVO grammar to overtake a V2 grammar, it is required that the SVO grammar has a greater *advantage*. That is, there must be more sentences in the linguistic evidence that are incompatible with the V2 grammar than with the SVO grammar. (71) shows the advantage patterns of V2 over SVO, and vice versa:[5]

(71) a. *Advantage of V2 grammar over SVO grammar*:

V2 $\rightarrow$ s but SVO $\not\rightarrow$ s: VS (XVSO, OVS).

b. *Advantage of SVO grammar over V2 grammar*:

SVO $\rightarrow$ s but V2 $\not\rightarrow$ s: V>2 (SXVO, XSVO).

[5]Again, for simplicity, we follow Lightfoot (1991) to consider only degree-0 sentences as linguistic input, although nothing hinges on this assumption.

If the distribution patterns in modern V2 languages are indicative of those of ancient times, we can see that the V2 constraint is in general very resilient to erosion. In languages like German, the V2 constraint is very strongly manifested. Matrix V>2 patterns are restricted to a small number of adverbs and other specific lexical items, and are quite rare in distribution:

(72) *Rare V>2 patterns in modern German*:

.... denn Johann hat gestern das Buch gelesen.
.... so Johann had yesterday the book read.

Statistical analysis of Dutch, German, Norwegian, and Swedish (cited in Lightfoot 1997) shows that about 70% of all sentences in V2 languages are SVO, and about 30% are VS patterns, which include XVSO and OVS. Our own counts based on a Dutch sample of adult-to-child speech (MacWhinney & Snow 1985) are similar: 66.8% SVO, 23% XVSO, and 1.2% OVS. In contrast, based on the Penn Treebank, a corpus of modern English, we found that only less than 10% of all sentences have V>2 word order:

(73) *V>2 patterns in modern English*:

a. He always reads newspapers in the morning.

b. Every night after dinner Charles and Emma played backgammon.

Therefore, the 10% advantage of SVO grammar, expressed in V>2 patterns, cannot throw off a V2 grammar, which has 30% of VS patterns to counter.

If the V2 constraint is so resilient, how on earth did Old French lose it? The reason, on our view, is that OF was also a null subject language.

Recall that the advantage of V2 grammar over SVO grammar is expressed in VS patterns. However, this advantage would be considerably diminished if the subject is dropped to yield [X V *pro*] patterns: a null subject SVO grammar (like modern Italian) can analyze such patterns as [X (*pro*) V]. (74) shows the prevalence of subject drop in early Middle French (MidFr):

(74)

Text	SV	VS	NullS	
Froissart, *Chroniques* (c. 1390)	40%	18%	42%	
15 Joyes (14esme Joye) (c. 1400)	52.5%	5%	42.5%	
Chartier *Quadrilogue* (1422)	51%	7%	42%	(R: p155)

The 30% advantage in non-*pro*-drop V2 languages has been reduced to 5-18% in the *pro*-drop MidFr. As the same time, V>2 patterns have gone from fairly sparse (about <5%) in OF (R: p95) to 11-15% in early MidFr, as the class of sentence-initial XPs that do not trigger SV inversion was expanded (Vance, 1989). (75) shows some representative examples:

(75) *V>2 patterns in early MidFr*:

a. Lors la royne *fist* Santré appeller.
then the queen made Santré to-call.
'Then the queen had Saintré called.'

b. Et a ce parolles le roy *demanda* quelz prieres ilz faisonient
And at these words the king asked what requests they made.

c. Apres disner le chevalier me *dist* ...
after dinner the knight to-me said ...
'After dinner the knight said to me ...'

(76), which is based the examination of the three texts in (74), shows the frequency of V>2 patterns in MidFr:

(76)

Text	V>2	
Froissart, *Chroniques* (c. 1390)	12% (of 73)	
15 Joyes (14esme Joye) (c. 1400)	15% (of 40)	
Chartier *Quadrilogue* (1422)	11% (of 45)	(R: p148)

Comparing (76) with (74), we see that at early MidFr stage, there were more V>2 sentences than VS sentences, due to the effect of subject drop. Thus, following the corollary in (68), it must be the case that an SVO grammar (plus *pro*-drop) has an advantage over an OF V2 grammar (plus *pro*-drop). V2 in French was then destined to extinction, as predicted.

Our analysis of the loss of V2 in French crucially relies on the fact that null subject was lost *after* V2 was lost. R shows that this was indeed the case. In late 15th century and early 16th century, when SVO orders had already become "favored", there was still significant use of null subjects, as the statistics in (77) demonstrate:

(77) *The lasting effect of pro-drop in MidFr*:

	SV	VS	NullS	
Anon., *Cent Nouvelles Nouvelles* (1466)	60.2%	10%	12%	
Anon., *Le Roman de Jehan de Paris* (1495)	60%	10%	30%	
Vigneulles, CNN (1505-15)	60%	11%	29%	(R: p155 and p199)

Overall, the mean figures for the relevant patterns are shown below (R: p199):

(78)

	SV	VS	NullS
15th century	48%	10%	42%
16th century	77%	3%	15%

The decline, and eventually, disappearance, of VS patterns are the result of the SVO grammar winning over the V2 grammar. We see that in the 16th century, when V2 almost completely evaporated, there was still considerable amount of subject drop. This diachronic pattern is consistent with our explanation for the loss of V2 in Old French.

We believe that the present analysis may be extended to other Western European Romance languages, which, as is well known, all had V2 in the medieval times. Under the present model of grammar competition, it is no accident that all such languages at one time had *pro*-drop, as in Old French, and many still do, as in Italian, Spanish, etc. It appears that the combination of *pro*-drop and V2 are intrinsically unstable, and will necessarily give away to a SVO (plus *pro*-drop) grammar. Without concrete statistics from the history of these languages, we can only extrapolate from their modern forms. It is reported (Bates, 1976) that modern Italian employs *pro*-drop in 70% of all sentences; as a result, the 30% advantage of a V2 grammar over an SVO grammar (in VS sentences) would be reduced to $30\% \times 30\% = 9\%$. Now this is a figure already lower than the approximately 10% of $V > 2$ sentences that an SVO grammar has an advantage over a V2 grammar, which would lead to the demise of V2.

5.3 The Erosion of V2 in Middle English

We now turn to the erosion of V2 in Middle English. Unless specified otherwise, all our examples and statistics are taken from Kroch & Taylor (1997, henceforth K&T). Our interpretation of the historical facts supports and formalizes their analysis.

5.3.1 Word order in Old English

K&T shows that Old English (OE) is, generally speaking, a Germanic language similar to Yiddish and Icelandic. Its peculiarities lie in the distribution of its V2 patterns, which are different from modern West Germanic languages such as Dutch and German (van Kemenade 1987, Pintzuk, 1991, K&T).

In OE, when the subject is an NP, the finite verb is in the second position:

(79) *V2 with NP subjects in OE*:

a. þæt hus hæfdon Romane to ðæm anum tacne geworht
that building had Romans with the one feature constructed.

b. þær wearþ se cyning Bagsecg ofslægen
there was the king Bagsecq slain

In contrast, a pronominal subject precedes the verb, creating superficially V3 patterns with a non-subject topic phrase:

(80) *V3 with pronoun subjects in OE*:

a. Ælc yfel he mæg don.
each evil he can do.

b. scortlice ic hæbbe nu gesæd ymb þa þrie dælas ...
briefly I have now spoken about the three parts.

c. ðfter his gebede he ahof þæt cild up ...
after his prayer he lifted the child up

The subject pronoun is often analyzed as a clitic (van Kemenade 1987, Pintzuk 1991).

Furthermore, there are genuine V3 patterns when the topic position is occupied by a certain class of temporal adverbs and adjuncts. In these constructions, the subject, pronominal or phrasal, precedes the verb: inversion:[6]

(81) *V3 with XP topics in OE*:

a. Her Oswald se eadiga arcebisceop forlet þis lif.
in-this-year Oswald the blessed archbishop forsook this life

[6]Although genuine V3 patterns are also possible in modern West Germanics, they are restricted to "it-then" sentences and left-dislocation such as (72). Their distributions are not as wide as in OE; see K&T for details.

b. On þisum geare Willelm cyng geaf Raulfe eorle Willelmes dohtor
In this year William king gave (to) Ralph earl William's daughter
Osbearnes sunu.
Osborn's son.

The V2 constraint is uniformly obeyed in questions, where the verb raises to C, the subject, be it pronoun or NP, is in the post-verbal position:

(82) *Verb raising to C in OE*:

a. hwi sceole we oþres mannes niman?
why should we another man's take

b. þa ge-mette he sceaðan.
then met he robbers

c. ne mihton hi nænigne fultum æt him begitan.
not could they not-any help from him get

d. hæfdon hi hiora onfangen ær Hæsten to Beamfleote come.
had they them received before Hæsten to Benfleet came

5.3.2 The southern dialect

K&T shows that there was considerable dialectical variation with respect to the V2 constraint in the period of early Middle English (ME),. Specifically, the southern dialect essentially preserved the V2 of Old English: proposed XPs, with exception of a a certain class of adverbs and adjuncts noted earlier, generally trigger subject-verb inversion with full NP subjects but rarely with pronoun subjects. Table (5.3.2), taken from Kroch, Taylor, & Ringe (1997: table 1), illustrates:

Following van Kemenade (1987), we relate the eventual loss of V2 in English to the loss of subject cliticization. The loss of subject cliticization (and that of word order freedom in general) can further be linked to impoverishment of the morphological case system of pronouns; see Kiparsky (1997) for a possible theoretical formulation of this traditional idea. Recall the V3 patterns in the southern dialect of early ME, which are manifested in sentences with pronominal subjects (80) and certain adverb and adjunct topics (81), schematically shown as in (83):

	NP subjects	Pronoun subjects
Preposed XP	% inverted	% inverted
NP complements	93% (50/54)	5% (4/88)
PP complements	75% (12/16)	0% (0/11)
Adj. complements	95% (20/21)	33% (7/21)
þ*a/then*	95% (37/39)	72% (26/36)
now	92% (12/13)	27% (8/30)
PP adjuncts	75% (56/75)	2% (2/101)
adverbs	57% (79/138)	1% (1/182)

Table 5.1: V2 in southern early Middle English.

(83) XP subject-pronoun V_{FIN} ...

With the impoverishment and eventual loss of the morphological case system, clitics are no longer possible. Therefore, patterns such as (83) were no longer compatible with an OE type V2 grammar. However, they *were* compatible with an SVO grammar with the subject-pronoun treated a DP, as in modern English. Examining Table 1, we can see that 62% (511/825) of all matrix sentences are of the V>2 pattern of the pattern (83) and 38% (314/825) are of the VS pattern. When subject pronoun could not be analyzed as clitics any more but only as NPs, the SVO grammar would have had a greater advantage than the V2 grammar, and eventually rose to dominance. The loss of morphological case system makes the loss of V2 possible, and the competition between the SVO grammar and the OE V2 grammar is straightforwardly captured in the present model of language change.

Notice that we also have an immediate account for the so-called "residual V2" in modern English questions, certain negations, etc. Recall that in (82), we have seen that when V raises to C, both pronoun and NP subjects are in post-verbal position. In other words, the linguistic evidence *for those constructions* has been homogeneous with respective to a V2 grammar throughout the history of English. Therefore, their V2 character is preserved.[7]

[7]More precisely, what has been preserved are the parametric choices that OE made in dimensions such as question and negation which the so-called "residual V2" are attributed to.

5.3.3 The northern dialect and language contact

In contrast to the southern dialect, K&T shows that the northern dialect, under heavy Scandinavian influence, was very much like modern Germanic languages. The V2 constraint was uniformly and rigidly enforced, and one does not find the almost categorical asymmetry between pronoun and NP subjects in Old English and southern early Middle English.

As noted earlier, the V2 constraint exhibited in West Germanic languages is difficult to overthrow. This is due to the advantage a V2 grammar has over competing grammars, e.g. an SVO grammar: V2 grammar generates VS sentences which punish SVO grammar, SVO grammar generates V>2 sentences which punish V2 grammar, but VS sentences usually outnumber V>2 sentences. In discussing the loss of V2 in Old French, we argued that subject drop in Old French considerably diminished V2's advantage, to a point where an SVO grammar, aided by an increase in V>2 patterns, eventually won out. How did the northern early Middle English, a rigid V2 language *without* subject drop, evolve into an SVO language?

K&T shows that the extensive contact between the northern and southern populations in the period of Middle English was essential to the eventual loss of V2 in English. They insightfully attribute the erosion of V2 to the competition of grammars in learners during language contact. This analysis is naturally formulated in the present model of language change. The northern V2 dialect, when mixed with the southern (essentially OE) language, constituted a heterogeneous linguistic environment for later generations of learners, who, instead of converging to a single grammar, attained a mixture of co-existing grammars. Table (5.3.3), taken from Kroch et al. (1997), shows the consequences of language contact in the northern dialect.

The effect of language contact is clear. Recall that prior to contact, the Northern dialect was much like Germanic languages in which V2 is strongly enforced: Kroch, et al. (1997) found subject verb inversion in 93.3% of all sentences containing subjects. After contact, shown in Table (5.3.3), while NP subjects still in general follow subjects, the overall subject verb inversion rate has dropped to 68.2% (308/305). This indicates that as a result of language contact and mixing, the V2 constraint in the Northern dialect was considerably weakened. When the V2 constraint is sufficiently weakened, and if the morphological case

	NP subjects	Pronoun subjects
Preposed XP	% inverted	% inverted
NP complements	100% (8/8)	64% (16/25)
PP complements	88% (21/24)	70% (48/69)
Adj. complements	100% (10/10)	25% (2/8)
then	86% (6/7)	51% (24/47)
now	100% (4/4)	82% (14/17)
adverbs	80% (20/25)	57% (35/61)

Table 5.2: V2 (after language contact) in the Northern ms. (Thornton) of the Mirror of St. Edmund.

system of the mixed language got lost, then an SVO grammar would have gradually taken over, in the manner described earlier for the loss of V2 in OE.

For the northern dialect, the initial contact with the southern dialect was crucial in the loss of V2.[8] That is, a West Germanic V2 language similar to the northern dialect, would not lose V2 without language contact, which introduces a substantial amount of V>2 patterns for the learner, even if its morphological case were lost. Mainland Scandinavian languages such as Swedish and Danish, with impoverished morphological case system but nevertheless strongly V2, presumably falls into this category. Once language contact was made, the homogeneity of linguistic evidence was punctured, which resulted in co-existence of two distinct grammars internalized in the learners. The loss of morphological case system resulted in the loss of the clitics system, which further favored the SVO grammar and eventually drove it to complete dominance.

5.4 Conclusion

We now summarize this preliminary investigation of an acquisition-based model of language change. Our approach is again motivated by Darwinian variational thinking, and is founded on two factual observations: (a) the deviation of child language from adult language is not simply noise or imperfection – it is the reflection of actual grammar hypotheses, as argued

[8] According to our theory, contact was not crucial for the Southern dialect to lose V2. The $V > 2$ patterns in the Southern dialect, which resulted from the lack of pronoun-verb inversion, would have gradually eliminated V2 after the clitic system was lost.

in the preceding chapters, and (b) the inherent variability of language use, and in particular, evidence of multiple grammars in mature speakers during the course of language change, as shown particularly clearly in Kroch and his colleagues' work.

The model formalizes historical linguists' intuition of grammar competition, and directly relates the statistical properties of historical texts (i.e., acquisition evidence) to the direction of language change. It is important to recognize that while sociological and other external forces clearly affect the composition of linguistic evidence, grammar competition as language acquisition, the locus of language change, is internal to the individual learner's mind/brain. We hope that the present model, by directly linking the statistical properties of historical text and the predictable outcome of language acquisition, will contribute to a formal framework in which problems in language change can be studied quantitatively.

Chapter 6

A Synthesis of Linguistic Knowledge and Learning

To end this preliminary study of the variational approach to language, let's return to the abstract formulation of language acquisition to situate the variational model in a broader context of cognitive studies.

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

The variational model a calls for a balanced view of S_0 and L: domain-specific knowledge of language as innate UG as well as domain-neutral theories of learning. UG allows the learner to go beyond unanalyzed distributional properties of the input data. The connection between S_0 and L is made possible by the variational and probabilistic thinking that are central to Darwinian evolutionary theory. Under variational thinking, children's deviation from adult language becomes the reflection of possible variations in human language; under probabilistic thinking, the continuous changes in the distributional patterns of child language are associated with discrete grammars of human language and their statistical distributions. The present approach, if correct, shows that a synthesis of Universal Grammar and psychological learning is not only possible, but also desirable.

We stress again that the admission of general learning into language acquisition in no way diminishes the importance of innate Universal Grammar in the understanding of natural language. Recall, for example (section 3.3), the presence of Chinese type topic-drop during

English children's Null Subject stage, as demonstrated by the almost categorical asymmetry in argument vs. adjunct NS questions (41) and the almost perfect match between Chinese and English children's NS/NO ratio (Table 3.3.2). It is inconceivable that such patterns can be explained without appealing to a very domain-specific property of human grammar. In fact, the search for evidence is entirely guided by linguistic theories developed on the basis of adult grammatical judgment: the typology of three grammars, English, Chinese, and Italian, and their associated syntactic properties, without which I wouldn't even know where to look.

On the other side of the coin, the variational model complements linguistics theories in a novel and interesting way: it provides an independent and theory-neutral tool for accessing their psychological status. A linguistic theory is an abstract description of language, which categorizes linguistic phenomenon into insightful and interesting ways: parameters, for example, are one of the devices used to capture Important generalizations and natural classes. Just as there is an infinitely many ways to slice up a cake, each of which a potential controversy, disagreement arises when what one linguist's insight to divide up the linguistic space is not shared by another.

If two competing linguistic theories, T_1 and T_2, are not merely restatement of each other (as is often the case), then they must capture different linguistic generalizations by making use of, say, two different parameters, P_1 and P_2, respectively. Suppose further that both of them give a descriptively adequate account of some range of linguistic data. If these descriptions of linguistic competence are to have any direct bearing on linguistic performance such as acquisition,[1], the differences between P_1 and P_2 will be manifested in the acquisition of their respective parameter values in the target grammar, once we plug the theory of S_0 into a theory-neutral model of learning $\mathcal{L}$, the variational model, different developmental consequences (D_1 and D_2) will presumably result.

(85) $T_1 \longrightarrow \boxed{\mathcal{L}} \longrightarrow D_1$
$T_2 \longrightarrow \boxed{\mathcal{L}} \longrightarrow D_2$

[1]A desirable feature of a competence theory but by no means a necessary one: see Yang (1996) for discussion in relation to the so-called "psychological reality" of linguistic theories.

(85) can be carried out straightforwardly: the identification of the relevant evidence to set target values, and the estimation of their frequencies in naturalistic corpora. Several aspects of D_1 and D_2 can then be evaluated in an acquisition text: developmental time course as compared to empirically established baselines (section 3.1 and 3.2), co-existence of competing parameter values (section 3.3), regularities in performance data (chapter 4), and, when available, diachronic trends through time (chapter 5). All things being equal, theories more compatible with facts from acquisition can be regarded as more plausible theories of Universal Grammar.

Following this line of reasoning, we can already reach the verdict on a number of contentious issues in linguistic theorizing. For example, the evidence presented in section 3.3 strongly suggests that the nature of subject drop should be understood as a bifurcation into the Italian, agreement-based type and the Chinese, discourse identification type, and that the obligatoriness of overt subject is associated with the presence of pure expletives (e.g., *there*) in the language. Alternative formulation of subject use, though descriptively perfectly adequate, are likely to be suspect. Similarly, the demonstration that English irregular verbs are organized in classes, defined by independent suffixation and readjustment rules, seems to vindicate a derivational conception of morphophonology.[2]

Variational thinking and statistical modeling proved instrumental in the theory of population genetics: they make a direct link between idealized and discrete Mendelian genetics and the variable patterns of biological evolution and diversity, which were apparently at odds. By the use of variational thinking and statistical modeling, it is now possible to relate directly generative linguists' idealized and discrete grammars to the variable patterns of human linguistic behavior. Perhaps, just perhaps, the variational approach provides a principled way of bridging or at least shortening a similar gap, which lies between linguistic competence and linguistic performance.

After a couple of Scotch on a rainy afternoon, I'd like to think that's the case.

[2]To the best of my knowledge, there has been no systematic treatment in a constraint-based Optimality Theory (Prince and Smolensky 1993) of what traditionally falls under the lexical phonology of English verbs (Halle and Mohanan 1985). Another topic in phonological acquisition that may bear on the derivation-constraint debate is the acquisition of stress, where a considerable amount of statistics on children's performance is available (Kehoe and Stoel-Gammon 1997, Demuth and Roark 1999); unfortunately, the most explicit OT treatment of stress (Tesar 1998) is essentially a restatement of the rule-based approach (Hayes 1994, Dresher 1999).

Chapter 7

Bibliography

Atkinson, Richard, Gorden Bower, and Edward Crothers. 1965. *An introduction to mathematical learning theory*. New York, New York: Wiley.

Bailey, C.-J. 1973. *Variation and linguistic theory*. Washington, DC: Center for Applied Linguistics.

Barton, Andrew, and Richard Sutton. 1998. *Reinforcement learning*. Cambridge, Mass.: MIT Press.

Bates, Elizabeth. 1976. *Language and context: the acquisition of pragmatics*. New York, New York: Academic Press.

Bates, Elizabeth, and Jeffrey Elman. 1996. Learning rediscovered: a perspective on Saffran, Aslin, and Newport. *Science* 274: 1849-1850.

Battye, Adrian and Ian Roberts. 1995. Introduction. In Adrian Battye and Ian Roberts (eds.). *Clause structure and language change*. New York: Oxford University Press.

Behrens, Heike. 1993. Temporal reference in German child language. Doctoral dissertation, University of Amsterdam, Amsterdam, the Netherlands.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14, 150-177.

Berwick, Robert. 1985. *The acquisition of syntactic knowledge*. Cambridge, Mass.: MIT Press.

Berwick, Robert, and Partha Niyogi. 1996. Learning from triggers. *Linguistic Inquiry* 27: 605-622.

Berwick, Robert, and Amy Weinberg. 1984. *The grammatical basis of linguistic performance.* Cambridge, Mass.: MIT Press.

Bloom, Louise. 1970. *Language development: Form and function in emerging grammars.* Cambridge, Mass.: MIT Press.

Bloom, Paul. 1990. Subjectless sentences in child languages. *Linguistic Inquiry* 21: 491-504.

Bloom, Paul. 1993. Grammatical continuity in language development: the case of subjectless sentences. *Linguistic Inquiry* 24: 721-34.

Bloomfield, Leonard. (1927). Review of J. O. H. Jespersen, *The philosophy of grammar. Journal of English and Germanic Philology* 26, 244-6.

Borer, Hagit, and Kenneth Wexler. 1987. The maturation of syntax. In *Parameter setting*, ed. Thomas Roeper and Edwin Williams. Dordrecht: Reidel.

Brown, Roger. 1973. *A first language.* Cambridge, Mass.: Harvard University Press.

Brown, Roger, and Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In Hayes, J. R. (ed.) *Cognition and the development of language.* New York: Wiley. 155-207.

Bush, Robert, and Frederic Mosteller. 1951. A mathematical model for simple learning. *Psychological Review* 68: 313-323.

Bush, Robert, and Frederic Mosteller. 1958. *Stochastic models for learning.* New York, New York: Wiley.

Caselli, M.C., E. Bates, Paula Casadio, J. Fenson, L. Fenson, L. Sanderl, and J. Weir. 1995. A cross-linguistic study of early lexical development. *Cognitive Development* 10: 159-199.

Chomsky, Noam. 1955. *The logical structure of linguistic theory.* Manuscript, Harvard and MIT. Published in 1975 by Plenum, New York.

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1975. *Reflections on language.* New York, New York: Pantheon.

Chomsky, Noam. 1977. On Wh-movement. In *Formal syntax*, eds. Peter Culicover, Thomas Wasow, and Adrian Akmajian. New York, New York: Academic Press.

Chomsky, Noam. 1981. *Lectures on government and binding.* Dordrecht, the Netherlands: Foris.

Chomsky, Noam. 1986. *Knowledge of language: its nature, origin, and use.* New York: Praeger.

Chomsky, Noam. 1995a. Language and nature. *Mind* 104: 1-61.

Chomsky, Noam. 1995b. *The minimalist program.* Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1999. Linguistics and brain sciences. Ms. MIT.

Chomsky, Noam, and Morris Halle. 1968. *The sound patterns of English.* Cambridge, MA: MIT Press.

Clahsen, Harald. 1986. Verbal inflections in German child language: acquisition of agreement markings and the functions they encode. *Linguistics* 24: 79-121.

Clahsen, Harald, and Martina Penke. 1992. The acquisition of agreement morphology and its syntactic consequences: new evidence on German child language from the Simone corpus. In *The acquisition of verb placement*, ed. Jürgen Meisel. Dordrecht, the Netherlands: Kluwer.

Clahsen, Harald, and M. Rothweiler. 1993. Inflectional rules in children's grammars: evidence from the development of participles in German. *Yearbook of Morphology 1992*, 1-34.

Clark, Robin. 1992. The selection of syntactic knowledge. *Language Acquisition* 2: 83-149.

Clark, Robin, and Ian Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24: 299-345.

Crain, Stephen. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences* 14: 597-650.

Crain, Stephen, and Mineharu Nakayama. 1987. Structure dependency in grammar formation. *Language* 63: 522-543.

Demuth, Katherine. 1989. Maturation and the acquisition of Sesotho passive. *Language* 65:56-90.

Demuth, Katherine, and Brian Roark. 1999. Talk given at the Boston University Conference on Language Development. Boston, MA.

Dresher, Bezalel Elan, and Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 34: 137-195.

Dresher, Bezalel Elan. 1999. Charting the learning path: cues to parameter setting. *Linguistic Inquiry* 30: 27-67.

Elman, Jeffrey. 1990. Finding structure in time. *Cognitive Science* 14, 179-211.

Elman, Jeffrey. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7, 195-224.

Fassi-Fehri, Abdelkader. 1993. *Issues in the structure of Arabic clauses and words.* Boston, Mass.: Kluwer.

Felix, Sasha. 1987. *Cognition and language growth.* Dordrecht: Kluwer.

Fodor, Janet Dean. 1998. Unambiguous triggers. *Linguistic Inquiry* 29: 1-36.

Fodor, Jerry, and Zenon Pylyshyn. 1988. Connectionist and cognitive architecture: a critical analysis. *Cognition* 28: 3-71.

Fong, Sandiway. 1991. The computational properties of principle-based grammatical theories. Ph.D. dissertation, MIT.

Francis, Nelson, and Henry Kucěra. 1982. *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Frank, Robert, and Shyam Kapur. 1996. On the use of triggers in parameter setting. *Linguistic Inquiry* 27: 623-660.

Gallistel, C. R. 1990. *The organization of learning.* Cambridge, Mass.: MIT Press.

Gerken, Lou Anne. 1991. The metrical basis for children's subjectless sentences. *Journal of Memory and Language* 30: 431-51.

Gibson, Edward and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25: 355-407.

Gleitman, Lila. 1981. Maturational determinants of language growth. *Cognition* 10: 115-126.

Gold, E. Mark. 1967. Language identification in the limit. *Information and Control*, 10: 447-74.

Guasti, Marie Teresa. 1992. Verb syntax in Italian child grammar. *Geneva Generative Papers* 1: 115-122.

Haegeman, Liliane. 1995. Root infinitives, tense, and truncated structures. *Language Acquisition* 4: 205-55.

Halle, Morris. 1990. An approach to morphology. *Proceedings of the Northeast Linguistic Society*, 20, 150-84.

Halle, Morris. 1997. Distributed Morphology: Impoverishment and fission. In *PF: Papers at the Interface.* Cambridge, MA: MITWPL. 425-450.

Halle, Morris, and K.-P. Mohanan, K. 1985. Segmental phonology of modern English. *Linguistic Inquiry* 16, 57-116.

Halle, Morris, and Alec Marantz. 1993. Distributed morphology. In *The view from Building 20*, eds. Hale, Kenneth and Samuel Jay Keyser. Cambridge, MA: MIT Press.

Hayes, Bruce. 1994. Metrical stress theory. Chicago, IL: University of Chicago Press.

Hockett, Charles. 1968. *The state of the art.* Mouton: the Hague.

Huang, James C.-H. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15: 531-574.

Hyams, Nina. 1986 *Language acquisition and the theory of parameters.* Dordrecht, the Netherlands: Reidel.

Hyams, Nina. 1991. A reanalysis of null subjects in child language. In *Theoretical issues in language acquisition : continuity change in development*, eds. Jürgen Weissenborn, Helen Goodluck, and Thomas Roeper. Hillsdale, New Jersey: L. Erlbaum Associates.

Hyams, Nina. 1996. The underspecification of functional categories in early grammar. In *Generative perspectives on language acquisition*, ed. Harald Clahsen. Amsterdam, the Netherlands: John Benjamins.

Hyams, Nina, and Kenneth Wexler. 1993. On the grammatical basis of null subjects in child language. *Linguistic Inquiry* 24: 421-459.

Jaeggli, Osvaldo, and Kenneth Safir. 1989. *The null subject parameter*. Boston, Mass.: Kluwer.

Jakobson, Roman. 1941/1968. *Child language, aphasia and phonological universals*. The Hague: Mouton.

Jenkins, Lyle. 1999. *Biolinguistics*. Cambridge, UK: Cambridge University Press.

Jusczyk, P., Kemler Nelson, D., Hirsh-Pasek, K., Kennedy, L., Woodward, A., and Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology* 24, 252-293.

Kayne, Richard. (1994). *The antisymmetry of syntax*. MIT Press: Cambridge, Mass.

Keenan, Edward. (1998). The historical creation of reflexive pronouns in English. Talk given at MIT, Cambridge, MA.

Kehoe, Margaret, and Carol Stoel-Gammon. 1997. The acquisition of prosodic structure: An investigation of current accounts of children's prosodic development. *Language* 73, 113-144.

Keyser, Samuel Jay, Greg Carlson, and Thomas Roeper. 1977. Dring, drang, drung. Ms., UMass Amherst.

Kim, John, Gary Marcus, Steven Pinker, Michael Hollander, and Michele Coppola. 1994. Sensitivity of children's inflection to morphological structure. *Journal of Child Language* 21, 173-209.

Kiparsky, Paul. 1973. "Elsewhere" in phonology. In *A festschrift for Morris Halle*, (eds.) Stephen Anderson and Paul Kiparsky. New York: Holt, Rinehart, and Winston. 93-106.

Kiparsky, Paul. (1997). The rise of positional licensing. In van Kemenade, Ans and Nigel Vincent. (eds.)

van Kemenade, Ans. (1987). *Syntactic case and morphological case in the history of English*. Dordrecht: Foris.

van Kemenade, Ans and Nigel Vincent. (eds.) (1997). *Parameters and morphosyntactic change.* New York: Cambridge University Press.

Kohl, Karen. 1999. An analysis of finite parameter learning in linguistic spaces. S.M. thesis, MIT, Cambridge, Mass.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language variation and change* 1: 199-244.

Kroch, Anthony, and Ann Taylor. 1997. Verb movement in Old and Middle English: dialect variation and language contact. In *Parameters of morphosyntactic change*, eds. Anns van Kemenade and Nigel Vincent. Cambridge, UK: Cambridge University Press.

Kroch, Anthony, Ann Taylor, and Donald Ringe. (1997). The Middle English verb-second constraint: a case study in language contact and language change. Manuscript. University of Pennsylvania.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715-762.

Lewontin, Richard. 1983. The organism as the subject and object of evolution. *Scientia* 118: 65-82.

Lewontin, Richard. 1996. Population genetics. In Anthony Griffith, Jeffrey Miler, David Suzuki, Richard Lewontin, and William Gelbart. *An introduction to genetic analysis.* Sixth edition. San Francisco, CA: W. H. Freeman.

Lightfoot, David. 1979. *The principles of diachronic syntax.* Cambridge, UK: Cambridge University Press.

Lightfoot, David. 1991. *How to set parameters.* Cambridge, MA: MIT Press.

Lightfoot, David. 1997. Shifting triggers and diachronic reanalysis. In *Parameters of morphosyntactic change*, eds. Anns van Kemenade and Nigel Vincent.

Macken, M. (1980). The child's lexical representation: the "puzzel-puddle-pickle" evidence. *Journal of Linguistics* 16, 1-17.

Macnamara, John. 1982. *Names for things: a study of human learning.* Cambridge, Mass.: MIT Press.

MacWhinney, Brian, and Catherine Snow. 1985. The Child Language Date Exchange System. *Journal of Child Language* 12: 271-296.

Marcus, Gary. 1993. Negative evidence in language acquisition. *Cognition* 46, 53-85.

Marcus, Gary. 1998. On eliminative connectionism. Ms. University of Massachusetts.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology* 29, 189-256.

Marcus, Gary, Steven Pinker, Michael Ullman, Michele Hollander, John Rosen, and Fei Xu. 1992. *Overregularization in language acquisition. Monographs of the Society for Research in Child Development*, 57.

Maynard Smith, John. 1989. *Evolutionary genetics.* Oxford, UK: Oxford University Press.

Mayr, Ernst. 1963. *Animal species and evolution.* Cambridge, Mass.: Harvard University Press.

Mayr, Ernst. 1982. *The growth of biological thought: Diversity, evolution, and inheritance.* Cambridge, Mass.: Harvard University Press.

Mehler, Jacques. Peter Jusczyk, Lambertz, G., Halstead, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor to language acquisition in young infants. *Cognition* 29, 143-178.

Molnar, Ray. forthcoming. The computational learning of irregular verbs. MEng. thesis, MIT.

Myers, Scott. 1987. Vowel shortening in English. *Natural Language and Linguistic Theory* 5: 485-518.

Narendra, Kumpati, and Mandayam Thathachar. 1989. *Learning automata.* Englewood Cliffs, New Jersey: Prentice Hall.

Niyogi, Partha and Robert Berwick. (1995). The logical problem of language change. Memo. MIT Artificial Intelligence Laboratory. Cambridge, MA.

Niyogi, Partha, and Robert Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61: 161-193.

Norman, Frank. 1972. *Markov processes and learning models.* New York, New York: Academic Press.

Osherson, Daniel, Scott Weinstein, and Michael Stob. 1982. A note on formal learning theory. *Cognition* 11: 77-88.

Piattelli-Palmarini, Massimo. 1989. Evolution, selection, and cognition: from "learning" to parameter setting in biology and in the study of language. *Cognition* 31: 1-44.

Phillips, Colin. 1995. Syntax at age 2: Cross-linguistic differences. In *MIT working papers in linguistics* 26, 325-382. Department of Linguistics and Philosophy, MIT, Cambridge, Mass.

Pierce, Amy. 1989. On the emergence of syntax : a crosslinguistic study. Ph.D. dissertation, MIT, Cambridge, Mass.

Pinker, Steven. 1979. Formal models of language learning. *Cognition* 7: 217-83.

Pinker, Steven. 1984. *Language learnability and language development.* Cambridge, Mass.: Harvard University Press.

Pinker, Susan. 1995. Why the child holded the baby rabbit: a case study in language acquisition. In *An invitation to cognitive science: Language*, eds. Gleitman, L. and Liberman, M. Cambridge, MA: MIT Press.

Pinker, Steven, and Alan Prince. 1988. On language and connectionism: analysis of a parallel distributed model of language acquisition. *Cognition* 28: 73-193.

Pinker, Steven, and Alan Prince. 1994. Regular and irregular morphology and the psychological status of rules of grammar. In Susan D. Lima, Roberta Corrigan, and Gregory K. Iverson. eds. *The reality of linguistic rules.* Amsterdam: John Benjamins.

Pintzuk, Susan. 1991. Phrase structure in competition: variation and change in Old English word order. Doctoral dissertation, University of Pennsylvania, Philadelphia, Pennsylvania.

Poeppel, David, and Kenneth Wexler. 1993. The full competence hypothesis. *Language* 69: 1-33.

Prince, Alan, and Paul Smolensky. 1993. *Optimality theory: constraint interaction in generative grammar.* Ms. Rutgers University and University of Colorado.

Pullum, Geoffrey. 1996. Learnability, hyperlearning, and the Poverty of the Stimulus. Paper presented at the Parasession on Learnability, 22nd Annual Meeting of the Berkeley Linguistics Society, Berkeley, CA.

Quang Phac Dong. 1968/1971. English sentences with overt grammatical subject. In *Studies out in left field: Defamatory essays presented to James D. McCawley*, Arnold Zwicky, Peter Salus, Robert Binnick, and Anthony Vanek (eds.) Amsterdam: John Benjamins.

Randall, Janet. 1990. The catapult hypothesis: an approach to unlearning. In *Theoretical issues in language acquisition : continuity change in development*, eds. Jürgen Weissenborn, Helen Goodluck, and Thomas Roeper. Hillsdale, New Jersey: L. Erlbaum Associates.

Roberts, Ian. 1993. *Verbs and diachronic syntax: A comparative history of English and French.* Dordrecht: Kluwer.

Roberts, Ian. 1997. Directionality and word order change in the history of English. In van Kemenade and Vincent (eds.).

Roeper, Thomas. 1973. Theoretical implications of word order, topicalization, and inflection in German language acquisition. In Charles Ferguson and Dan Slobin (eds.) *Studies in child language development.* New York: Holt, Rinehart, and Winston.

Roeper, Thomas. 2000. Universal Bilingualism. *Bilingualism: Language and Cognition* 2.

Rosch, Eleanor. 1978. Principles of categorization: A historical view. In *Cognition and categorization*, (eds.) Rosch, E. and Lloyd, B. B. Mahwah, NJ: Erlbaum.

Rizzi, Luigi. 1986. Null object in Italian and the theory of pro. *Linguistic Inquiry* 17: 501-557.

Rizzi, Luigi. 1994. Some notes on linguistic theory and language development: the case of root infinitives. *Language Acquisition* 3: 371-393.

Rumelhart, D. and McClelland, J. (1986). On learning the past tenses of English verbs.

Implicit rules or parallel distributed processing? In *Parallel distributed processing: Explorations in the microstructure of cognition*, (eds.) McClelland, J., Rumelhart, D., and the PDP Research Group. Cambridge, MA: MIT Press.

Saffran, Jennifer, Richard Aslin, and Elissa Newport. 1996. Statistical learning by 8-month old infants. *Science* 274: 1926-1928.

Sankoff, David. ed. 1978. *Language variation: models and methods.* New York, New York: Academic Press.

Sampson, Geoffrey. 1989. Language acquisition: growth or learning? *Philosophical Papers* 18, 203-240.

Sano, Tetsuya, and Nina Hyams. 1994. Agreement, finiteness, and the development of null arguments. In *Proceedings of NELS 24.* GLSA, University of Massachusetts, Amherst.

Santorini, Beatrice. 1992. Variation and change in Yiddish subordinate clause word order. *Natural Language and Linguistic Theory* 10: 595-640.

Seidenberg, Mark. 1997. Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275: 1599-1604.

Smith, Neil. 1973. *The acquisition of phonology: A case study.* Cambridge: Cambridge University Press.

Stampe, David. 1979. *A dissertation on natural phonology*. New York, New York: Garland.

Stromswold, Karen. 1990. Learnability and the acquisition of auxiliaries. Doctoral dissertation, MIT, Cambridge, Mass.

Tesar, Bruce. 1998. . An Iterative Strategy for Language Learning. *Lingua* 104:131-145

Travis, Lisa. 1984. Parameters and effects of word order variation. Doctoral dissertation, MIT, Cambridge, Mass.

Ullman, M., Corkin, S., Pinker, S., Coppola, M., Locascio, J., and Growdon, J. H. (1993). Neural modularity in language: evidence from Alzheimer's and Parkinson's disease. (abstract). Paper presented at the 23rd annual meeting of the Society for Neuroscience. Washington DC. Also in *Journal of Cognitive Neuroscience* 9: 289-299.

Valian, Virginia. 1990. Null subjects: A problem for parameter-setting models of language acquisition. *Cognition* 35: 105-122.

Valian, Virginia. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40: 21-82.

Waller, Bradley. 1997. Against a metrical basis for subject drop in child language. Ms. MIT.

Wang, Qi., Diane Lillo-Martin, Catherine Best, and Andrea Levitt. 1992. Null subject vs. null object: Some evidence from the acquisition of Chinese and English. *Language Acquisition* 2: 221-54.

Weinberg, Amy. 1990. Markedness vs. maturation: the case of subject-auxiliary inversion. *Language Acquisition* 1: 169-194.

Weinreich, Uriel., William Labov, and M. I. Herzog (1968). Empirical foundations for a theory of language change. *Directions for historical linguistics: A symposium*, ed. W. Lehman and Y. Malkiel. Austin, Texas: University of Texas Press.

Weverink, Meika. 1989. The subject in relation to inflection in child language. M.A. thesis, University of Utrecht, Utrecht, the Netherlands.

Wexler, Kenneth. 1994. Optional infinitives, head movement, and the economy of derivation in child language. In *Verb movement*, eds. David Lightfoot and Norbert Hornstein. Cambridge, UK: Cambridge University Press.

Wexler, Kenneth. 1998. Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. *Lingua* 106: 23-79.

Wexler, Kenneth, and Peter Culicover. 1980. *Formal principles of language acquisition.* Cambridge, Mass.: MIT Press.

Wijnen, Frank. 1999. Verb placement in Dutch child language: A longitudinal analysis. Ms., University of Utrecht, Utrecht, the Netherlands.

Xu, Fei, and Steven Pinker. 1995. Weird past tense forms. *Journal of Child Language* 22, 531-556.

Yang, Charles D. 1995. A minimalist parser. The 17th Annual CUNY Sentence Processing Conference. CUNY, New York.

Yang, Charles D. 1996. Psychological reality, type transparency, and minimalist sentence processing. Ms. Massachusetts Institute of Technology.

Yang, Charles D. 1999. Two grammars are better than one. Commentary on Roeper's Universal Bilingualism. *Bilingualism: Language and Cognition*, 2.

Yang, Charles D. 2000. Dig-dug, think-thunk. Review of *Words and Rules* by Steven Pinker. *The London Review of Books*, Vol 22, No. 16.

Yang, Charles, and Sam Gutmann. 1999. Language learning via Martingales. The Sixth Conference of the Mathematics of Language. Orlando, FL.