A Non-Deterministic, Success-Driven Model of Parameter Setting in Language Acquisition

Ph.D. Oral Examination

April 28, 1992

Eric H. Nyberg, 3rd
Program in Computational Linguistics
Department of Philosophy
Carnegie Mellon University

Introduction

I. Overview

II. Model Definition

III. Learning Domains

IV. Empirical Results

V. Evaluation

VI. Alternative Models

VII. Contributions

VIII. Future Work

I. Overview

- Parameter Theory
- The Input Data
- The Subset Principle
- Computational Models
- Evaluation Criteria
- Previous Models
- Challenges

Parameter Theory

The *principles and parameters* framework:

- Grammar is represented as a set of principles, or constraints on the form of possible grammars
- Cross-linguistic variation is captured by parameterized differences in the definitions of principles

Example: Parameters in Metrical Phonology

- *Direction of Constituent Construction (P_0).* If P_0 is 0, then metrical feet are constructed from left to right; if P_0 is 1, feet are constructed from right to left.
- *Headedness of Constituents (P_1).* If P_1 is 0, then constituents (feet) are left-headed; if P_1 is 1, they are right-headed.
- *Boundedness of Constituents (P_2).* If P_2 is 0, then constituents (feet) are bounded (binary); if P_2 is 1, then constituents are unbounded in size.

(Dresher and Kaye, 1990)

Parameters in Metrical Phonology

P_0	P_1	P_2	Metrical Feet
0	0	0	((A* B) (C* D) (E*))
0	0	1	((A* B C D E))
0	1	0	((A B*) (C D*) (E*))
0	1	1	((A B C D E*))
1	0	0	((A*) (B* C) (D* E))
1	0	1	((A* B C D E))
1	1	0	((A*) (B C*) (D E*))
1	1	1	((A B C D E*))

- The goal of parameter-setting is to learn the right values for each of the parameters
- Realistic systems for metrical phonology have many more parameters ($\approx$ 10)

The Input Data

- Positive Evidence ($s \in L_{target}$)
- Negative Evidence ($s \notin L_{target}$)
- Indirect Negative Evidence (inferred)

The Subset Principle

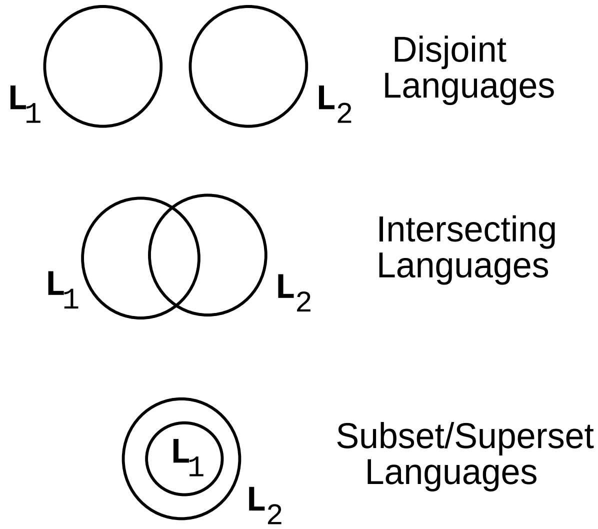


Figure 1: **Disjoint, Intersecting, and Subset/Superset Languages**

The Subset Principle. If there are two languages, $L_1 \subset L_2$ and the learner encounters data that are consistent with both languages, then it selects the smaller language, L_1.

(Angluin, 1978; Wexler and Culicover, 1980; Berwick 1985)

Computational Models

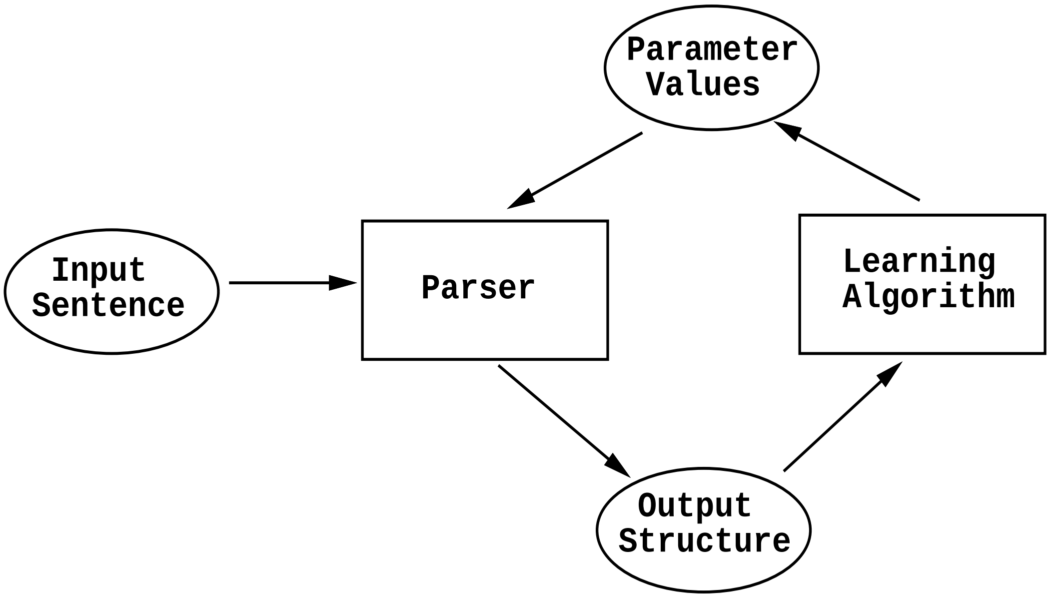


Figure 2: **Basic Parameter-Setting Architecture**

Formal Criteria

- The learning paradigm:
  1. Learner (learning algorithm)
  2. Search Space (set of languages to be learned)
  3. Learning Environment (set of possible data presentations)
  4. Learning Behavior (succession of hypotheses)
- Learning on infinite texts vs. finite texts
- Self-monitoring learners
- Horizon effects and formal properties in the limit

Computational Criteria

- Efficiency (Space and Time)
- Learning on finite texts
- Is getting the “right” answer good enough?

Cognitive Criteria

1. Learnability Condition
2. Equipotentiality Condition
3. Time Condition
4. Input Condition
5. Developmental Condition
6. Cognitive Condition

(Pinker, 1979)

YOUPIE (Dresher and Kaye)

- 11 parameters of metrical phonology
- Batch learning (perfect memory)
- Single hypothesis, deterministic learning
- Use of indirect negative evidence
- Learning behavior predetermined
- Cannot learn in the presence of noise

Single Hypothesis Learning

- Single-hypothesis, no-backtracking, error-driven systems:
  - A single active hypothesis h at any time t
  - Change the value of one parameter if $s_t \notin L(h)$
  - Cannot change the value of a parameter from 1 to 0
    (no backtracking)
- Parameter flipping implies susceptability to noise
- Single hypothesis + no backtracking makes it impossible to guarantee success on some learning problems (Hypothesis Ordering, Subset Shift)

Hypothesis Ordering, Subset Shift

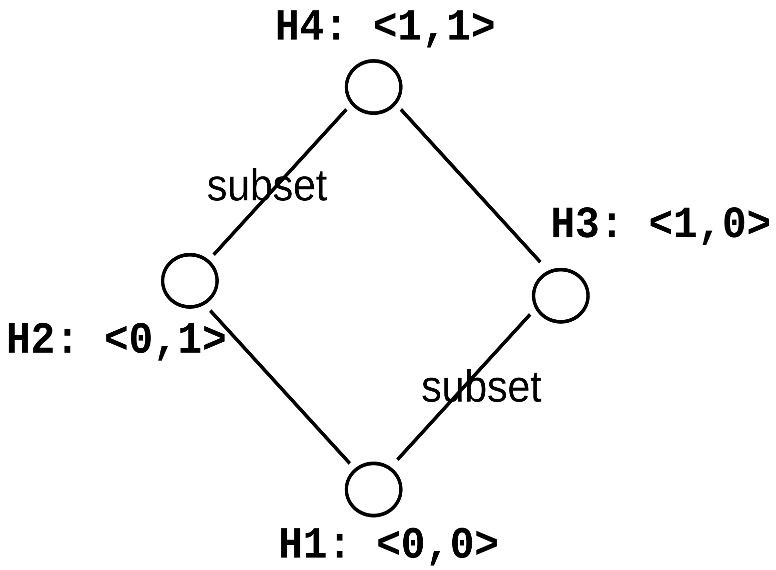


Figure 3: **A Simple Hypothesis Ordering Problem**

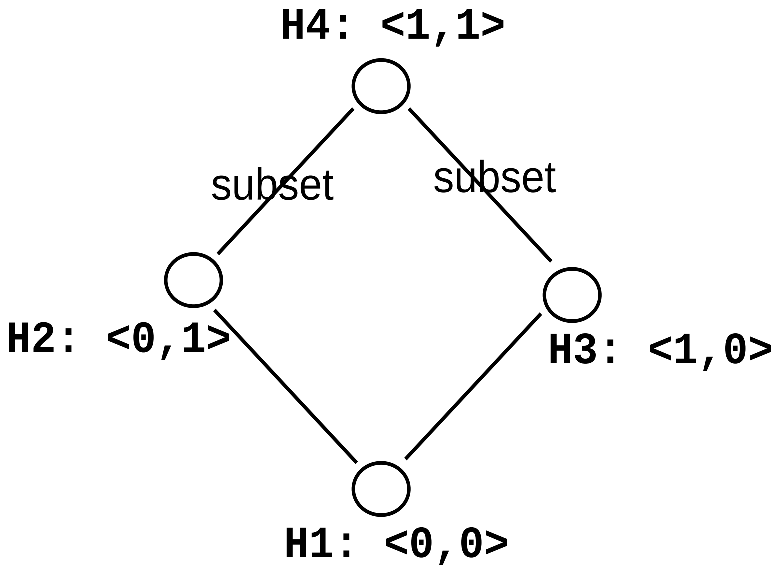


Figure 4: **The Subset Shift Problem**

DARWIN (Clark)

- Abstract syntactic parameters
- Genetic Algorithms approach
- Multiple hypotheses, non-deterministic learning
- No use of indirect negative evidence
- Can solve the Hypothesis Ordering and Subset Shift problems
- Hypotheses fluctuate widely
- Non-determinism is unconstrained
- Not tested in realistic environments

Challenges

- Present a complete model (parameter theory, learning algorithm, computational simulation and empirical results)
- No reliance on negative evidence
- Robust when learning in noisy environments
- Not prone to hypothesis ordering problems
- Limited non-determinism
- Conservative hypothesis selection
- Feedback to linguistic theory

II. Model Definition

- Hypotheses and Weighting
- Hypothesis Activation
- Learning Algorithm
- Avoiding Overgeneralization
- Example

Hypotheses and Weighting

- The set $DATA_t$ is the set of example sentences encountered by the learner up to time t;
- The set $POS_t(h) = s \in L(h), s \in DATA_t$;
- The set $NEG_t(h) = s \notin L(h), s \in DATA_t$;
- The value of $NPE_t(h) = |POS_t(h)| - |NEG_t(h)|$.

The weight associated with each active hypothesis is a function of the NPE, calculated by a sigmoid function:

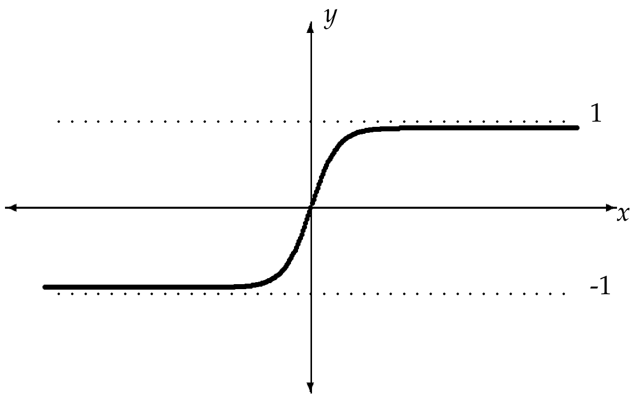


Figure 5: **The sigmoid function $y = 2/(1 + e^{-x}) - 1$.**

Biased Weighting

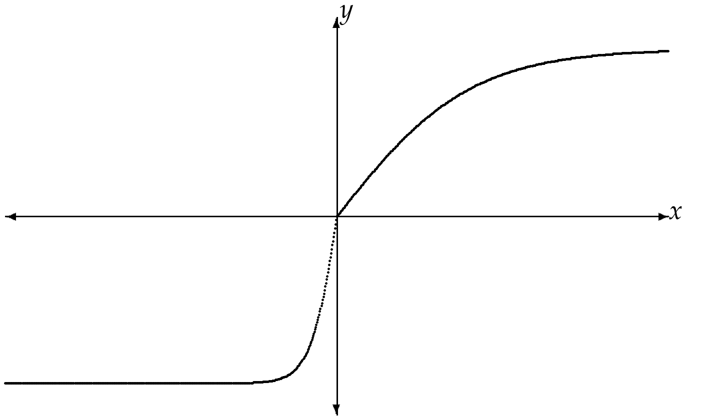


Figure 6: Sigmoid with Biased Weighting

Hypothesis Activation

- *The Single-Value Constraint*. At any time t, if the learner holds some hypothesis h, then it may activate only hypotheses that differ from h by the value of a single parameter.
- The Hypothesis Graph:

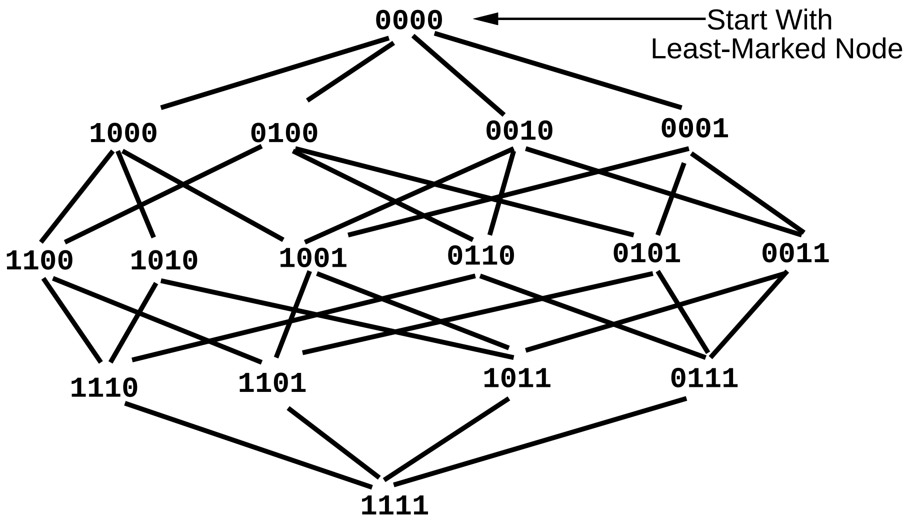


Figure 7: **1-Adjacent Hypothesis Graph,** $n = 4$

- How can the graph be searched without generating all the nodes?

Learning Algorithm

1. Start with the unmarked hypothesis (all 0's) as
2. Loop: for each active hypothesis h,
   - Read in a piece of data s
   - If $s \in L(h)$, then add increment the weight to s, else decrement weight (according to sigmoid function)
   - If the weight of h dips below $(\epsilon - 1)$, then prune h
3. If no hypotheses are active, then activate the neighbors of the last node that was pruned.

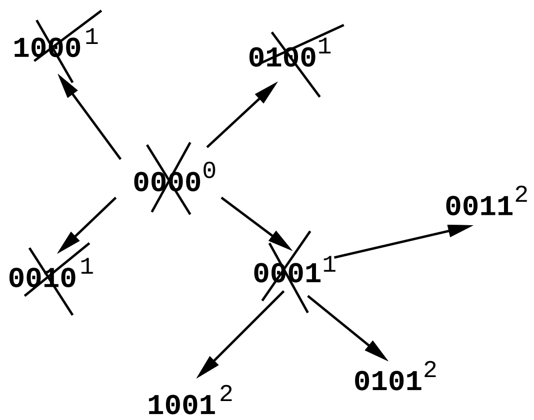


Figure 8: **Hypothesis Activation**

Avoiding Overgeneralization

- If the weight of h becomes greater than $(1 - \epsilon)$, then activate any subset languages that haven't been activated, and
- Deactivate any superset languages that are still active.

Example

III. Learning Domains

- Metrical Phonology
- Syntax
- Noisy Environments

Metrical Phonology

- Eleven parameters from (Dresher and Kaye, 1990)
- 432 languages in the search space
- Full computational implementation of a metrical stress assigner
- Integration with the learning algorithm

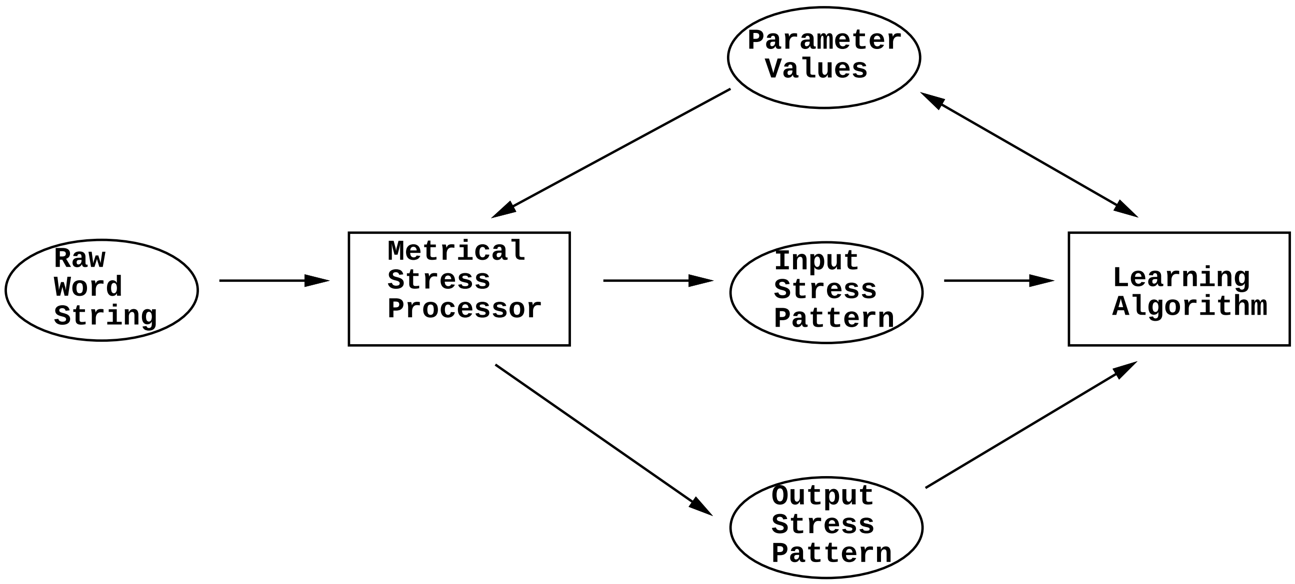
Metrical Parameters

P0	The word-tree is strong on the [Left/Right]
P1	Feet are [Binary/Unbounded]
P2	Feet are built from the [Left/Right]
P3	Feet are strong on the [Left/Right]
P4	Feet are quantity sensitive (QS) [No/Yes]
P5	Feet are QS to the [Rime/Nucleus]
P6	A strong branch of a foot must itself branch [No/Yes]
P7	There is an extrametrical syllable [No/Yes]
P8	It is extrametrical on the [Left/Right]
P9	A weak foot is defooted in clash [No/Yes]
P10	Feet are noniterative [No/Yes]

Stress Assignment

- Syllable construction
- Foot construction
- Secondary stress assignment
- Primary stress assignment
- Stress adjustment

Integration with the Learner



Syntax

- Abstract syntactic parameters (Clark, 1990)
- Test generality of approach (different domain, varying size of search space, varying expressiveness of data, etc.)
- Realistic models probably contain > 20 parameters

Noisy Environments

- Abstract syntactic parameters (Clark, 1990)
- Test robustness of algorithm in noisy learning environments
- Random noise injector model (sporadic vs. burst noise)
- Tested on noise rates 1/20, 1/10, 1/8, 1/6

IV. Results

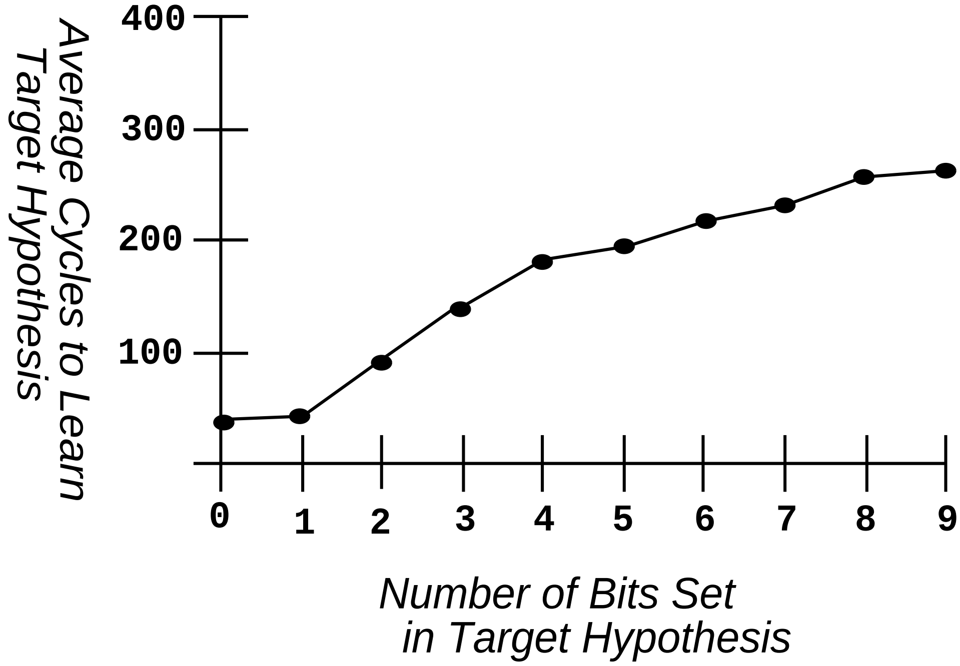
- Metrical Phonology
- Syntax
- Noisy Environments

Phonology Results

No. Bits Set	No. Hypotheses in Space	Avg. No. Examples
0	1	31
1	7	40
2	24	93
3	54	146
4	87	180
5	102	198
6	86	215
7	50	228
8	18	251
9	3	256
10	0	—
11	0	—

Figure 9: **Average Number of Input Examples by Number of Bits Set**

Phonology Results



Identically-Stressed Languages

```
 A: 00010000010

Metrical constituents (RIGHT-headed, BINARY feet built from the LEFT):
Final output (weak feet are defooted in clash):

     [  0        2 +    ][  0        1      ][  0 -    ]
       / \      / \        / \      / \        / \
      K   *    K   *      K   *    K   *      K   *
         / \      / \        / \      / \        / \
        O        OO         OO  K    O   K      OO  K
```

```
B: 00000001000

Metrical constituents (LEFT-headed, BINARY feet built from the LEFT)
  (the LEFT-most segment is extrametrical):
Final output:

        0       [  2 +      0      ][  1        0      ]
       / \        / \      / \        / \      / \
      K   *      K   *    K   *      K   *    K   *
         / \        / \      / \        / \      / \
        O          OO       OO  K      O   K    OO  K
```

Figure 10: **Languages with Identical Stress Patterns**

Highly Overlapping Languages

Word Length	Identical Stress	Different Stress
1	4	0
2	16	0
3	48	16
4	224	32
5	960	64
6	3968	128
7	16128	256

Figure 11: **Overlap Between Languages 01011101010 and 01001101010**

Syntax Results

Bits Set	Total Bits	Expr.	Avg. Inputs	Bits Set	Total Bits	Expr.	Avg. Inputs
2	4	1	141	10	20	4	438
2	4	2	77	10	20	8	326
2	4	3	63	10	20	12	180
2	4	4	52	10	20	16	172
4	4	1	179	10	20	20	186
4	4	2	98	20	20	4	501
4	4	3	76	20	20	8	383
4	4	4	70	20	20	12	270
5	10	2	353	20	20	16	253
5	10	4	191	20	20	20	254
5	10	6	117	15	30	6	500
5	10	8	98	15	30	12	367
5	10	10	87	15	30	18	265
10	10	2	489	15	30	24	278
10	10	4	217	15	30	30	325
10	10	6	153	30	30	6	588
10	10	8	133	30	30	12	493
10	10	10	125	30	30	18	406
				30	30	24	412
				30	30	30	430

Figure 12: **Learning 4, 10, 20 and 30 Parameters**

Syntax Results

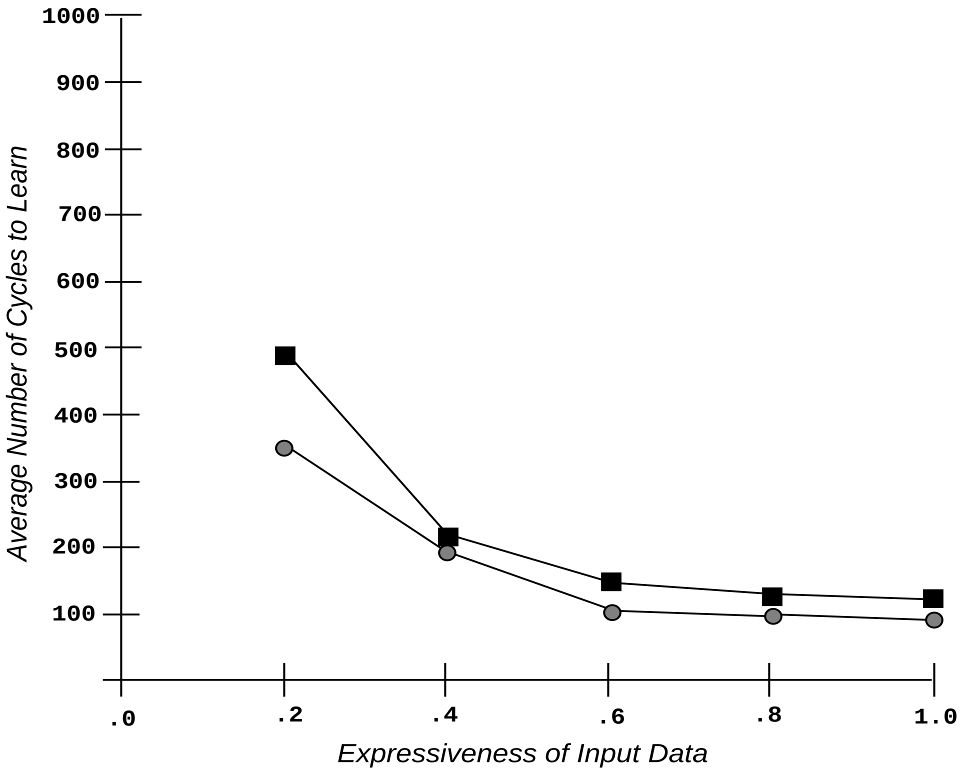


Figure 13: **Average Number of Inputs,** $n = 10$

Noise Results

- Theoretical limit on noise: 16.6
- Uniform random injection of noise into text
- Random error rates: 1/20 (5%), 1/10 (10%), 1/8 (12.5%), 1/6 (16.6%)

Noise Results

Bits Set	Total Bits	Expr.	1/20 Error	1/10 Error	1/8 Error	1/6 Error
2	4	1	119	160	204	273
2	4	2	89	85	151	142
2	4	3	73	82	171	281
2	4	4	66	122	158	199
4	4	1	157	171	174	452
4	4	2	107	142	182	192
4	4	3	88	120	164	319
4	4	4	79	121	142	243
5	10	2	429	289	288	382
5	10	4	164	193	190	226
5	10	6	122	125	213	442
5	10	8	114	121	148	229
5	10	10	98	120	219	357
10	10	2	451	302	341	317
10	10	4	197	216	240	291
10	10	6	167	185	223	251
10	10	8	163	196	181	442
10	10	10	135	194	207	226
10	20	4	546	397	409	400
10	20	8	264	228	338	415
10	20	12	185	239	244	NA
10	20	16	170	210	294	NA
10	20	20	190	246	267	NA
20	20	4	591	497	514	NA
20	20	8	332	340	385	NA
20	20	12	263	304	370	NA
20	20	16	253	274	362	NA
20	20	20	294	327	365	NA

Learning with Noise Bursts

- Model not implemented to withstand "bursts" of noisy examples
- Not encountered empirically
- Can occur (with low probability) given noise injector technique
- Straightforward extensions to the model possible

Learning with Noise Bursts

- Learning with Infinite Data Presentations.
  Problem: Learner may prune the correct hypothesis on infinitely many noise bursts
  Solution: Maintain memory of highest previous activation
- Avoiding Overgeneralization.
  Problem: A timely noise burst may prune the correct subset hypothesis during hypothesis "tightening"
  Solutions: Weighting subset hypotheses equally, Periodic subset activation

V. Evaluation

- Problems Solved
- Outstanding Problems
- Cognitive Criteria

Problems Solved

- Hypothesis ordering problems
- Robustness under uniform noise (up to 16.6%)
- Limited non-determinism (multiple hypotheses)
- Conservative hypothesis activation (learning behavior)
- Testbed for evaluation-by-simulation of parameter models

Outstanding Problems

- Learning on infinite texts with arbitrarily long noise bursts
- Testing against a large set of psycholinguistic data
- Full integration with a linguistic performance model

Cognitive Criteria

1. Learnability Condition
2. Equipotentiality Condition
3. Time Condition
4. Input Condition
5. Developmental Condition
6. Cognitive Condition

VI. Alternative Models

- Bayesian Models
  - Independence
  - Cognitive Plausibility
  - Conditional Probabilities
- Connectionist Models
  - Naturalness
  - Complexity

VII. Contributions

- Complete model
- No reliance on negative evidence
- Can learn in noisy environments
- Non-determinism solves hypothesis ordering problems
- Limited non-determinism more cognitively plausible
- Conservative hypothesis selection more cognitively plausible
- Empirical results give feedback to metrical parameter theory
- Alternative learning paradigm

VIII. Future Work

1. Short-Range Investigation.
   - Learning on infinite texts with noise bursts
   - Avoiding overgeneralization with noise bursts
   - Experimenting with Algorithm Coefficients
   - Further refinement of metrical parameters
   - Testing the learner on noisy data in the metrical domain
2. Long-Range Investigation.
   - Empirical data and stages in phonological development (massive psycholinguistic data gathering)
   - A complete syntactic processing model
   - Noise model for natural language
   - Frequency model for natural language
   - Convergence model for natural language