NECESSARY BIAS IN NATURAL LANGUAGE LEARNING

By

Lisa Sue Pearl

Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, in partial fulfillment of the requirements for the degree of Doctor of Philosophy
2007

Advisory Committee:
Professor Amy Weinberg, Chair
Professor Jeffrey Lidz
Professor William Idsardi
Professor Charles Yang
Professor James Reggia
Dedications

To Amy Weinberg, who put up with a ridiculous number of last minute questions and drafts of *everything*. With good humor, to boot.

To Norbert Hornstein, who was always exceptional at convincing me that my ideas were any good. He’s really quite persuasive.

To Charles Yang, whose work inspired all of this. More than once.

To Jeff Lidz, who helped me figure out how to write things that were actually comprehensible and who was always full of even more ideas.

To Bill Idsardi, who was also always full of even more ideas and who made me think very carefully about all my mathematical work.

To Philip Resnik, whose boundless energy and enthusiasm and questions completely inspired me time and time again.

To David Lightfoot, who taught the first theoretical linguistics class I ever took and who I wanted to grow up to be.

To Michelle Hugue, who taught me how to make computation into a tool and about how there are so many different ways to measure success.

To Peggy Antonisse, whose thoroughly sensible advice and calming influence were welcome balm in times of great stress and panic.
Acknowledgements

This work would have been a long time in coming without the listening ears, unwavering support, and good advice of many, many people, among them: Amy Weinberg, Norbert Hornstein, Charles Yang, Jeff Lidz, Bill Idsardi, Philip Resnik, David Lightfoot, Michelle Hugue, Peggy Antonisse, Colin Phillips, Andrea Zukowski, Heather Taylor, Brian Dillon, and Ivano Caponigro. This work was additionally supported by a National Science Foundation Graduate Research Fellowship.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedications</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>Chapter 1: A Theory of the Language Learning Mechanism</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The Mechanism of Language Learning</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Language Development: Constraints on the Hypothesis Space</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Formalizing the Language Acquisition Mechanism</td>
<td>2</td>
</tr>
<tr>
<td>1.4 Domain Specificity and Domain Generality</td>
<td>2</td>
</tr>
<tr>
<td>1.5 Investigating the Components of the Learning Framework</td>
<td>3</td>
</tr>
<tr>
<td>1.6 Computational Investigations of Data Intake Filtering</td>
<td>4</td>
</tr>
<tr>
<td>1.7 Organization of Dissertation</td>
<td>4</td>
</tr>
<tr>
<td>Chapter 2: Bayesian Updating in a Linguistic Framework</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Bayesian Updating: Overview</td>
<td>6</td>
</tr>
<tr>
<td>2.1.1 A Simple Case: Two Non-overlapping Hypotheses, Equally Likely</td>
<td>6</td>
</tr>
<tr>
<td>2.1.2. A Variant on the Simple Case: Two Non-overlapping Hypotheses, with an Initial Bias for One Hypothesis</td>
<td>8</td>
</tr>
<tr>
<td>2.1.3 A Less Simple Case: Two Overlapping Hypotheses, Equally Likely</td>
<td>9</td>
</tr>
<tr>
<td>2.1.4 A Variant of the Less Simple Case: Two Overlapping Hypotheses, with an Initial Bias for One Hypothesis</td>
<td>11</td>
</tr>
<tr>
<td>2.1.5 An Even Less Simple Case: Two Overlapping Hypotheses in a Subset Relation, Equally Likely</td>
<td>11</td>
</tr>
<tr>
<td>2.1.6 Hypothesis Spaces for Language Learning</td>
<td>16</td>
</tr>
<tr>
<td>2.2 Bayesian Updating: General Implementation for Language Learning in a Hypothesis Space with Two Hypotheses</td>
<td>17</td>
</tr>
<tr>
<td>2.2.1 Updating with Unambiguous Data in a Hypothesis Space with Two Hypotheses</td>
<td>17</td>
</tr>
<tr>
<td>2.2.2 Updating with Ambiguous Data in a Hypothesis Space with Two Hypotheses</td>
<td>21</td>
</tr>
<tr>
<td>2.2.3 About t</td>
<td>22</td>
</tr>
<tr>
<td>2.3 Summary of Bayesian Updating Adapted to a Linguistic Framework</td>
<td>25</td>
</tr>
<tr>
<td>Chapter 3: The Case of Anaphoric One</td>
<td>26</td>
</tr>
<tr>
<td>3.1 Anaphoric One: The Necessity of Domain-Specific Constraints</td>
<td>26</td>
</tr>
<tr>
<td>3.2 Why Learning Anaphoric One Is Interesting</td>
<td>27</td>
</tr>
<tr>
<td>3.3 Anaphoric One</td>
<td>28</td>
</tr>
<tr>
<td>3.3.1 Adult Knowledge: Syntactic Categories and Semantic Referents</td>
<td>28</td>
</tr>
<tr>
<td>3.3.2 The Pragmatics of Anaphoric One</td>
<td>30</td>
</tr>
<tr>
<td>3.3.3 Children’s Knowledge of Anaphoric One</td>
<td>30</td>
</tr>
<tr>
<td>3.3.4. Sparse Data for Anaphoric One</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Learning Anaphoric One</td>
<td>34</td>
</tr>
</tbody>
</table>
5.7.2 The Logical Problem of Learning English Metrical Phonology .......... 129
5.3.1 The Algorithm for Identifying All Viable Parameter-Setting Orders..... 130
      5.7.3.2 Relativization of Unambiguous Data Probability.................... 131
      5.7.3.3 An Example of Testing a Parameter-Setting Order................... 133
5.8 English Learning Results .................................................................... 134
      5.8.1 Order Constraints as a Metric ................................................. 134
      5.8.2 Parameter-setting Orders that Lead to English Target Values ....... 134
      5.8.3 Deriving Constraints .................................................................. 137
          5.8.3.1 Cues Method with Relative-Against-All: Accounting for Constraints .................................................................................. 138
          5.8.3.2 Parsing Method with Relative-Against-All/Potential: Accounting for Constraints ................................................................. 139
5.9 Discussion ......................................................................................... 140
      5.9.1 Cues: Why Better Constraints on Parameter-Setting Order? ....... 140
      5.9.2 Relativization .............................................................................. 141
      5.9.3 Cues and Parsing: A Viable Combination? ................................. 141
5.10 Summary ......................................................................................... 146
Chapter 6: Learning By Filtering .............................................................. 148
Appendix ................................................................................................... 149
Bibliography ............................................................................................ 161
List of Tables

Table 2.1: The effect of data intake accumulation on parameter-setting ...............24
Table 3.1: The distribution of utterances in the corpus examined by LWF ..........32
Table 3.2: The expected distribution of utterances in the input to young learners ....35
Table 3.3: The expected distribution of utterances in the input to the Bayesian learner for updating the semantics hypothesis.........................................................50
Table 4.1: OV order advantage in the input for the D0 and D1 clauses...............85
Table 4.2: Percentage of ambiguous clauses in the historical corpora..............87
Table 4.3: Formalization of quantities available from historical corpora and quantities to derive .................................................................88
Table 4.4: Data quantities after normalization...................................................88
Table 4.5: Derived quantities rewritten ............................................................90
Table 4.6: Quantities available from historical corpora and quantities to derive ......93
Table 4.7: Data quantities after normalization...................................................93
Table 4.8: Data from historical corpora and calculate p\textsubscript{VO}..................94
Table 4.9: OV order advantage in the D0 clauses............................................97
Table 4.10: Data gathered from speech directed to young children..................100
Table 4.11: OV order advantage at 1000 A.D. with no filters..........................100
Table 5.1: A comparison of the properties of the cues and parsing methods.....111
Table 5.2: The results of using the cues and parsing methods on Maranungku ....129
Table 5.3: Relativize-against-all approach for cues .........................................131
Table 5.4: Relativize-against-potential approach for cues ...............................132
Table 5.5: Relativize-against potential for parsing .........................................133
Table 5.6: Initial unambiguous data distribution from metrical phonology corpus..133
Table 5.7: Revised unambiguous data distribution from corpus .......................134
Table 5.8: Comparison of success of methods and relativization approaches........135
Table 5.9: Comparing cues and parsing with different relativization approaches....140
List of Figures

Figure 1: Two non-overlapping hypotheses, equally probable initially...................... 7
Figure 2: Two non-overlapping hypotheses with equal initial probability after seeing various distributions of intake................................................................. 7
Figure 3: Two non-overlapping hypotheses, with an initial bias towards A ............. 8
Figure 4: Two non-overlapping hypotheses with an initial bias for A after seeing various distributions and quantities of intake......................................................... 9
Figure 5: Two overlapping hypotheses, with equal probability initially............... 10
Figure 6: Two overlapping hypotheses in a subset relation, with equal probability initially.......................................................... 11
Figure 7: Two overlapping hypotheses in a subset relation with equal probability initially, after seeing unambiguous data points for the superset ....................... 12
Figure 8: Two overlapping hypotheses in a subset relation with equal probability initially, after seeing more subset data points.............................................. 13
Figure 9: Two overlapping hypotheses in a subset relation with equal probability initially, after seeing a mixed distribution............................................... 14
Figure 10: The Naive Parameter Learner, learning from subset data only........... 15
Figure 11: Structures for N' strings this ball and this red ball........................... 29
Figure 12: LWF experimental setup................................................................. 30
Figure 13: The set of potential referents for one for an utterance such as I have a red ball, and Jack has one, too ................................................................. 37
Figure 14: The observed set of referents for one for an utterance such as I have a red ball, and Jack has one, too ................................................................. 37
Figure 15: Comparison of different ratios of superset to subset, the likelihood of choosing a member of ther subset, and the effect on subset bias..................... 38
Figure 16: The syntax hypothesis space ............................................................ 40
Figure 17: The semantic hypothesis space ...................................................... 41
Figure 18: Linkings across domains................................................................. 42
Figure 19: Double update procedure, part 1................................................. 53
Figure 20: Double update procedure, part 2................................................. 54
Figure 21: Type II ambiguous update procedure........................................... 55
Figure 22: Decision tree for anaphoric one interpretation................................ 56
Figure 23: Idealized learner trajectory ......................................................... 56
Figure 24: EO Bayesian Learner's trajectory..................................................... 58
Figure 25: Varying t ..................................................................................... 59
Figure 26: Varying t, with incomplete learning period..................................... 59
Figure 27: Bayesian learning trajectory with filters....................................... 64
Figure 28: Cues as underspecified pieces of structure ....................................... 72
Figure 29: Degree-0 unambiguous learning trajectory.................................... 96
Figure 30: Learning trajectory with ambiguous data .................................... 97
Figure 31: Learning with degree-1 data ......................................................... 99
Figure 32: Metrical phonology parameters.................................................... 113
Figure 33: Sample metrical phonology structure........................................... 114