Experience-Based Language Acquisition:
A Computational Model of Human Language Acquisition

Brian E. Pangburn

ISBN: 1-58112-171-7

DISSERTATION.COM

USA • 2002

Experience-Based Language Acquisition:
A Computational Model of Human Language Acquisition

Dissertation.com
USA • 2002

ISBN: 1-58112-171-7

www.dissertation.com/library/1121717a.htm

To Jack and Jeremiah

Modeling, it should be clear, is an art form.
It depends on the experience and taste of the modeler.
- John Holland, *Hidden Order*

ACKNOWLEDGEMENTS

I would like to begin by thanking my wife, Jaimee, for standing behind me and for believing that I would one day finish my dissertation. Without her sacrifice, this dissertation would not be. I would like to thank my family for instilling in me the importance of a good education and express my gratitude to everyone at The Pangburn Company for being flexible and allowing me the time that I needed for research. I would also like to thank my committee for their patience as life, time and again, interrupted my research.

The material in this work on language acquisition in children is a direct result of my firsthand exposure to the brilliance of both Emily Smith and Jan Norris. Many of the technical roadblocks in the EBLA system were overcome thanks to the insights and friendship of Chris Branton and Carey Patin. The readability of this document owes much to the grammatical prowess of Sara Aguillard, who has always had the patience to proofread my work. Finally, there is no way that this research would have been finished in time without the tremendous help of Jon Ayo, who aided me with the EDISON port and did much of the filming and editing on the seemingly endless pile of videos for EBLA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Almost from the very beginning of the digital age, people have sought better ways to communicate with computers. This research investigates how computers might be enabled to *understand* natural language in a more humanlike way. Based, in part, on cognitive development in infants, we introduce an open computational framework for visual perception *and* grounded language acquisition called Experience-Based Language Acquisition (EBLA). EBLA can "watch" a series of short videos and acquire a simple language of nouns and verbs corresponding to the objects and object-object relations in those videos. Upon acquiring this *protolanguage*, EBLA can perform basic scene analysis to generate descriptions of novel videos.

The general architecture of EBLA is comprised of three stages: vision processing, entity extraction, and lexical resolution. In the vision processing stage, EBLA processes the individual frames in short videos, using a variation of the mean shift analysis image segmentation algorithm to identify and store information about significant objects. In the entity extraction stage, EBLA abstracts information about the significant objects in each video and the relationships among those objects into internal representations called *entities*. Finally, in the lexical acquisition stage, EBLA extracts the individual lexemes (words) from simple descriptions of each video and attempts to generate entity-lexeme mappings using an inference technique called cross-situational learning. EBLA is not primed with a base lexicon, so it faces the task of bootstrapping its lexicon from scratch.

The performance of EBLA has been evaluated based on acquisition speed and accuracy of scene descriptions. For a test set of simple animations, EBLA had average acquisition success rates as high as 100% and average description success rates as high as 96.7%. For a larger set

of real videos, EBLA had average acquisition success rates as high as 95.8% and average description success rates as high as 65.3%. The lower description success rate for the videos is attributed to the wide variance in entities across the videos.

While there have been several systems capable of learning object or event labels for videos, EBLA is the first known system to acquire both nouns *and* verbs using a grounded computer vision system.

CHAPTER 1 – INTRODUCTION

1.1 – A Problem of Overwhelming Difficulty

In a recent book about HAL, the computer in the movie *2001: A Space Odyssey*, David G. Stork wrote:

> Imagine, for example, a computer that could look at an arbitrary scene—anything from a sunset over a fishing village to Grand Central Station at rush hour—and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers. (Stork 2000, 8)

Unfortunately, true humanlike scene analysis is even more difficult than Stork indicates. This is because the solution to the language problem may very well depend on the solution to the vision problem, and on the broader problem of perception in general. Sensory perception gives meaning to much of human language, and to convey such meaning to a computer may require that perception be integrated with language from the very start.

The goal of this research is to construct a simplified version of the dynamic scene analysis system described by Stork (Stork 2000) and to investigate how computers might be enabled to understand language in more humanlike terms. While traditional, top-down research fields such as natural language processing (NLP), computational linguistics, and speech recognition and synthesis have made great progress in allowing computers to *process* natural language, they typically do not address *perceptual understanding*. In these fields, meaning and context for a given word are based solely on other words and the logical relationships among them.

To make this clearer, consider the following Webster's definition of *apple*: "The fleshy usually rounded and red or yellow edible pome fruit of a tree of the rose family." (Webster's 1989) Using traditional approaches, a computer might be able to determine from such a definition that an apple is "edible," that it is a "fruit," and that it is usually "rounded and red or yellow." But what does is *mean* to be "rounded and red"? People understand these words because their conceptual representations are grounded in their perceptual experiences. As for more abstract words, many have perceptual analogs or can be defined in terms of grounded words. Although it is unlikely that any two people share identical representations of a given word, there are generally enough similarities for that word to convey meaning. If computers can be enabled to ground language in perception, ultimately communication between man and machine may be facilitated.

1.2 – A Partial Solution

This research investigates the challenges of cognitive development and language acquisition for both children and computers. It details a new software framework, Experience-Based Language Acquisition (EBLA), that acquires a childlike language known as protolanguage in a bottom-up fashion based on visually perceived experiences. EBLA uses an integrated computer vision system to watch short videos and to generate internal representations of both the objects and the object-object relations in those videos. It then performs language acquisition by resolving these internal representations to the individual words in protolanguage descriptions of each video. Upon acquiring this grounded protolanguage, EBLA can perform basic scene analysis to generate simplistic descriptions of what it "sees."

EBLA operates in three primary stages: vision processing, entity extraction, and lexical resolution. In the vision processing stage, EBLA is presented with *experiences* in the form of short videos, each containing a simple event such as a hand picking up a ball. EBLA processes the individual frames in the videos to identify and store information about significant objects. In the entity extraction stage, EBLA aggregates the information from the video processing stage into internal representations called *entities*. Entities are defined for both the significant objects in each experience and for the relationships among those objects. Finally, in the lexical acquisition stage, EBLA attempts to acquire language for the entities extracted in the second stage using protolanguage descriptions of each event. It extracts the individual lexemes (words) from each description and then attempts to generate entity-lexeme mappings using an inference technique called cross-situational learning. EBLA is not primed with a base lexicon, so it faces the task of bootstrapping its lexicon from scratch.

For example, assume EBLA is presented with a short video of a hand picking up a ball and the protolanguage description "hand pickup ball." In the video processing phase, EBLA would extract the individual frames from the movie (see figure 1), and determine the location of the significant objects in each (see figure 2).

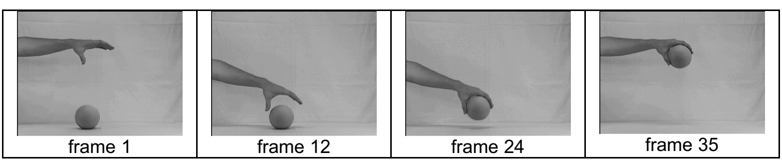

Figure 1. Frames from an Experience Processed by EBLA

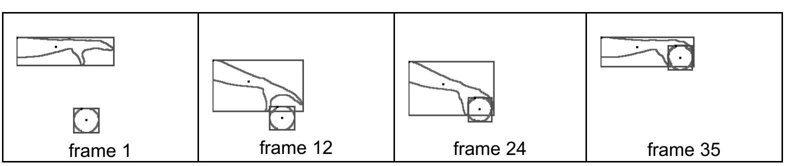

Figure 2. Frames Following the Detection of Significant Objects

In the entity extraction phase, EBLA would analyze the information in all of the frames to establish object entity definitions for the hand and the ball, and a relation entity definition for the spatial relationship between the hand and the ball. In the lexical resolution stage, EBLA would attempt to resolve the lexemes "hand," "pickup," and "ball" to their respective entities.

Since EBLA is not primed with any entities, lexemes, or mappings, it faces ambiguity in its early experiences (see figure 3). If the above example were its first experience, it would have no way to establish any of the entity-lexeme mappings. In order to overcome this, EBLA compares both entities and lexemes across multiple experiences to resolve ambiguity. A more detailed discussion of this process is presented in section 4.7.

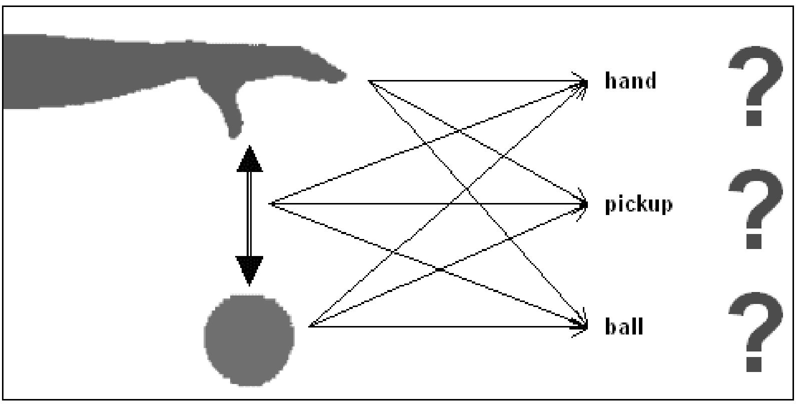

Figure 3. Referential Ambiguity Faced by EBLA

In order to implement the EBLA project in a reasonable amount of time using available technologies, the model has been constrained in several ways. First, EBLA's perceptual system is limited to a two-dimensional computer vision system. Second, the protolanguage descriptions delivered to EBLA are in textual form. A graphical representation of these first two abstractions is provided in figure 4. The third way that EBLA has been constrained is that it only attempts to acquire an unstructured protolanguage of nouns and verbs. EBLA cannot resolve other parts of speech such as adjectives or adverbs, and it makes no attempt to incorporate syntax. Finally, EBLA operates in an unsupervised manner in that it does not get any feedback on its performance. Hopefully, all of these constraints present a worst-case scenario, and by adding additional perceptual capabilities, language structure, or a feedback system, its performance could be improved. A more detailed description of the project constraints is presented in section 4.4.

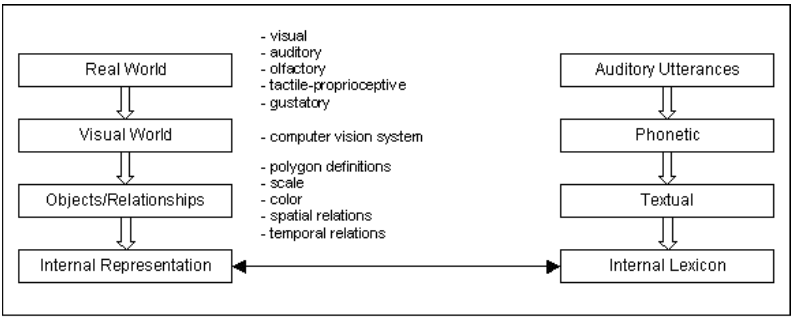

Figure 4. Abstractions for Computational Model

In order to facilitate the future elimination of some of the constraints placed on the EBLA Model, it has been designed as an open framework with expansion and extension in mind. For example, it should be a fairly straightforward process to modify EBLA to accommodate a speech recognition or a speech synthesis module. EBLA has been coded entirely in Java and documented following the JavaDoc conventions. The open-source PostgreSQL database system is used for all of EBLA's data storage and retrieval. EBLA has been developed using open-source and/or freely available tools whenever possible and has been successfully tested on both the Windows and Linux platforms. A listing of available resources for EBLA is provided in appendix B.

1.3 – What Lies Beneath

Chapter 2 investigates current theories of cognitive development and lexical acquisition in infants and toddlers in order to establish some developmental basis for EBLA. First, an overview of the nature versus nurture debate is provided, focusing on research that lies in the middle and attempts to bridge the two philosophies. Next, two integrated models of development are presented: the Experiential Model formulated by Katherine Nelson, a Professor of Developmental Psychology at City University of New York, and the Situational-Discourse-Semantic (SDS) Model developed by Janet Norris and Paul Hoffman, Professors of Communication Sciences and Disorders at Louisiana State University. Chapter 2 concludes with a chronological outline of cognitive and linguistic development through the first year-and-a-half of life based on the Experiential Model and the SDS Model as well as other pertinent research.

Chapter 3 provides a summary of several bottom-up computational models of grounded perception and language acquisition related to EBLA. These include a system that acquires word-to-meaning mappings for conceptual symbols; several related systems that perform vision-based event recognition; a system that acquires verbs using a virtual proprioceptive model; a system that acquires color and shape words using a vision system and recorded audio; a system that acquires object names, spatial terms, and syntax for computer generated images of various rectangles; and finally, a system that performs vision-based acquisition of object names based on social mediation.

Chapter 4 introduces the EBLA Model in general terms. It begins with some remarks about how the model relates to existing developmental and computational research. Next, the abstractions and constraints used for EBLA are outlined, and the experiences used to evaluate the model are discussed. The chapter concludes with an overview of the entity recognition and lexical acquisition mechanisms employed by EBLA.

Chapter 5 discusses the technology behind EBLA and details the model's implementation on a module-by-module level. Chapter 6 details the evaluation of EBLA including the methods used, the data sets involved, and the results obtained. Chapter 7 summaries long-term and short-term goals for further research and outlines the dissemination plans for EBLA. Finally, chapter 8 summarizes this research with a discussion about the significance of the results, a comparison with other research, and remarks on possible applications.

CHAPTER 2 – EARLY LANGUAGE ACQUISITION IN CHILDREN

2.1 – Introduction

This chapter surveys a variety of theories of early cognitive development and language acquisition in children. A better understanding of the processes involved in child development can provide much insight into how a computational model might accomplish the same tasks. This is not to say that the model detailed in this work is a simulation of human processes, but rather that the child and computer face analogous hurdles.

Research on child development involves a wide variety of studies including cognitive science, developmental psychology, communication disorders, linguistics, biology, and genetics. Unfortunately, there is no single unifying theory agreed to by each of the involved domains of research. Since the task of this research is *modeling*, it will only be necessary to extract common features among the theories that lend themselves to the development of a bottom-up computational model.

2.2 – Why Can't We All Just Get Along?

As an outsider to the field, it can be rather confusing to study the sciences behind child development. While the processes by which a newborn develops the skills to function as a member of society are fascinating, they are the subject of much debate. Much of this debate seems unnecessary as the various camps on development have more in common than one would be led to believe. At the root of the problem is the age-old argument of nature versus nurture. In the next few sections, some of the more common theories of development are discussed.

2.2.1 – Nature

The nature-centric view of development is known as nativism and focuses on functions and behaviors that are innate. Generally members of this camp believe that infants are born with innate, domain-specific cognitive structure to handle specialized functions including face recognition, language processing, and mathematics. Learning is little more than a process of fine-tuning this cognitive structure for a particular environment. For example, grammar, is thought to have universal principles that underlie all languages. As children are exposed to their native language, they tune parameters for that language, determining things such as verb placement in phrases. For further discussion of nativism, see Elman, et. al. (Elman, et. al.1999) and Pinker (Pinker 2000).

A variation of the nativist view of development is the evolutionist view. While some in the nativist camp believe that innate cognitive structures come into existence spontaneously as coherent functional units, in the evolutionist view, all innate function is seen as a direct result of Darwinian biological evolution. Language and other brain functions are thought to have evolved slowly over time, thus appearing in some crude form in the evolutionary ancestors of man. The distinction between the nativist and evolutionist views of development dates back to a conflict between Charles Darwin and linguist Max Muller in the late 1800's. Muller took the stance that language is one of the major traits separating humans from the rest of the animals and, therefore, could not have evolved from some related function in lower animals. Unfortunately, about a century later, modern linguist Noam Chomsky took a seemingly related stance regarding language that perpetuated the divide between the nativist and evolutionist views.

> His combination of an insistence on the biological nature of language with a refusal to look at the origins of that nature—and his blanket statements about the futility of any such enterprise—turned off many in the evolutionary community who might otherwise have been supportive. (Calvin and Bickerton 2001, 198)

2.2.2 – Nurture

The nurture-centric view of development is known as empiricism and focuses on environmental effects on learning. Generally, members of this camp believe that infants are born with only domain-general cognitive structure and learn specialized functions based on environmental stimuli. Infants are thought to be born without any task-specific modules. "Learning, in this view, involves a copying or internalizing of behaviors which are present in the environment." (Elman, et. al. 1999, 1) The behaviorism movement in psychology, which focuses on stimulus-response mechanisms as a basis for learning, is one of the more well-known empiricist approaches.

2.2.3 – Epigenesis

Many of the more modern theories of development are based on epigenesis, a *combination* of nature and nurture. The roots of epigenesis lie in the works of Piaget and other classic developmentalists during the first half of the twentieth century. They began to study both genetic *and* environmental factors as the path to cognition. (Nelson 1998) Generally, members of the epigenetic camp believe that domain-specific cognitive structure *emerges* from the *interactions* between domain-general cognitive structure and experience.

A more recent technical spin on epigenesis is the popular connectionist view of development. It combines discoveries about the workings of the brain with modern computational techniques to model various cognitive processes. (Elman, et. al. 1999)

As extensions to the epigenetic viewpoint, several newer models have integrated the specific impact of social and cultural mediation on child development. Two of these models actually form the developmental basis for this research and are discussed in detail later in this chapter.

2.2.4 – Everything in Moderation

The problem with much of the literature on language and development is that researchers far too often entrench themselves in one camp or another and then quote the competition out of context to prove a point. Linguists are portrayed as studying language in a vacuum, naively believing in a magical "language organ" or "grammar gene." Connectionists are portrayed as oversimplifying language and building toy models that only achieve limited results. All of the camps are stereotyped, and these stereotypes are often based on antiquated themes.

Fortunately, several recent works have finally undertaken the tasks of dispelling myths and attempting to reconcile the disparate camps on language and development. These works bring Chomsky back in line with Darwin, and demonstrate how a connectionist network might produce innate function if genetics control much of the wiring. (Calvin and Bickerton 2001; Pinker 2000) The truth is that no one yet completely understands the brain, language, *or* child development. Prejudices aside, most modern researchers can gain insights from certain principles of nativism/evolutionism, empiricism, *and* epigenesis.

Evolution today implies a lot more than it did in the days of Darwin and Muller. Natural selection is only part of the picture. Genetics have shown that while certain genes can be tied to specific traits, it is the complex interaction among large sets of genes that make humans *uniquely human*. Concepts such as autocatalytic sets and complexity theory have provided plausible explanations for nonlinear, emergent behavior in evolution. (Kauffman 1993; 1995; Waldrop 1993) As researchers continue to discover more about the human genome, there is no doubt that science will reveal what is and is not innate. Recently, in fact, it was discovered that mutations in the FOXP2 gene about 200,000 years ago may have given humans the capacity for speech.

> A mounting body of research suggests that the mutant gene conferred on human ancestors a finer degree of control over muscles of the mouth and throat, possibly giving those ancestors a rich new palette of sounds that could serve as the foundation of language. (Gillis 2002)

While there are few pure empiricists in modern times, one cannot simply dismiss the fact that there are a lot of environment-dependent concepts that humans learn. The world changes far too quickly for many types of behavior and function to be innate. Knowledge of how to program a personal computer, for example, is in no way innate. Humans must adapt to an ever-changing world using a skill set quite different from that of their ancestors.

At first glance, epigenesis seems to achieve a happy medium between nature and nurture, but studying the interactions of genetics and environment is a big undertaking and has a long way to go. Connectionism seems to be a promising avenue, but if not applied carefully, it can easily be reclassified as a form of empiricism. The neural networks most commonly used to model connectionism are powerful tools capable of learning by capturing complex, nonlinear relationships among stimuli, but they generally learn from experience. Sometimes the only innate, "genetic" components of a neural network are the underlying learning algorithm and the assumptions and constraints on the model. (Elman, et. al. 1999)

For all three camps, the biggest debate seems to be the extent to which cognitive function, and in particular, language, is innate. Modern linguists such as Pinker seem to believe that there are innate circuits for language in the brain, but that there is no single "language organ."

> The developing brain may take advantage of the disembodied nature of computation to position language circuits with some degree of flexibility. Say a variety of brain areas have the potential to grow the precise wiring diagrams for language components. An initial bias causes the circuits to be laid down