EBLA: A Perceptually Grounded Model of Language Acquisition

Brian E. Pangburn
The Pangburn Company, Inc.
103 Gisele Street
P.O. Box 900
New Roads, LA 70760-0900
bpangburn@nqadmin.com

S. Sitharama Iyengar
Department of Computer Science
298 Coates Hall
Louisiana State University
Baton Rouge, LA 70803
iyengar@bit.csc.lsu.edu

Robert C. Mathews
Department of Psychology
236 Audubon Hall
Louisiana State University
Baton Rouge, LA 70803
psmath@lsu.edu

Jonathan P. Ayo
Department of Information Systems
and Decision Sciences
Louisiana State University
8001 Jefferson Hwy., Apt. 99
Baton Rouge, LA 70809
jonayo2@hotmail.com

Abstract

This paper introduces 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 performance of EBLA has been evaluated based on accuracy and speed of protolanguage acquisition as well as on accuracy of generated 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 the appearance of objects across the test set.

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.

1 Introduction

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.

This paper 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.

While, to date, EBLA has only been evaluated using short descriptions comprised of nouns and verbs, one of the primary goals of this research has been to develop an open system that can potentially learn any perceptually grounded lexeme using a unified approach. The entities recognized EBLA are generic in nature and are comprised of clusters of perceptual attributes linked in a database system. Although only twelve basic attributes have been programmed into the current system, both the EBLA software and database support the addition of other attributes. There are even mechanisms in the database to support dynamic loading/unloading of custom attribute calculators.

2 Related Work

EBLA is based on research into language acquisition in children as well as existing computational models. This section highlights some of this related research. For a more detailed discussion of existing works on early language acquisition in children including works by Calvin and Bickerton (2001), Lakoff (1990), Locke (1993), Norris and Hoffman (2002), Pinker (2000), and Smith (1999), see chapter 2 of Pangburn (2002). For a more detailed discussion of existing computational models including Steels and Kaplan (2000) and Roy (1999; 2000), see chapter 3 of Pangburn (2002).

2.1 Experiential Model of Child Development and Language Acquisition

Katherine Nelson (1998) has worked to bring together many of the domains involved in the cognitive development of children with special emphasis on the role played by language. She views language and cognition as heavily intertwined—language cannot develop without early, nonlinguistic cognitive function, and full cognitive development cannot occur without language. Nelson takes an *experiential* approach to her work, focusing on how children adapt to meet their current needs and how that adaptation then affects their future experiences.

Nelson's Experiential Model is centered on *events* in the child's environment rather than *objects*. Nelson broadly defines an event as "an organized sequence of actions through time and space that has a perceived goal or end point." (Nelson 1998, 93-94) Events place objects and actions on those objects in the context of their ultimate goal or purpose, adding temporal ordering with a beginning and an ending. A child's perception, processing, classification, and storage of events form his/her *mental event representations* (MERs). The MER becomes the cognitive building block for increasingly complex knowledge representation and, ultimately, natural language.

2.2 Cross-Situational Techniques for Lexical Acquisition

Throughout the 1990's, Siskind (1992; 1997) has established algorithms to map words to symbolic representations of their meanings. For example, given the utterance, "*John walked to school.*" and a symbolic representation of the event, "GO(**John**, TO(**school**))," his system would learn the mappings, "*John*→**John**, *walked*→GO(x, y), *t* →TO(x), and *school*→**school**."

To perform the word-to-meaning mappings, Siskind utilizes cross-situational learning. Basically, this means that the system resolves mappings only after being presented with multiple utterance/symbolic concept sets representing multiple *situations*. By drawing inferences about word mappings from multiple uses, the system is able to determine the correct symbolic mappings.

2.3 Force Dynamics and Event Logic for Grounded Event Recognition

In distinct but related research, Siskind (1992; 2000; Siskind and Morris 1996) has developed several soft-

ware systems to classify and describe dynamic events. In 1992, he described ABIGAIL, a system that constructs semantic descriptions of events occurring in computer-generated stick-figure animations. ABIGAIL *perceives* events by detecting support, contact, and attachment using counterfactual simulation.

Using a subsequent system named HOWARD, Siskind and Morris built event representations based on real video. HOWARD produces hidden Markov models (HMMs) of the motion profiles of the objects involved in an event.

Siskind's most recent approach has been to use event-logic to describe changes in support, contact, and attachment, which he now terms *force-dynamics*. His latest system, LEONARD, uses a camera to capture a sequence of images and then processes that sequence using three subroutines:

1. Segmentation-and-Tracking – places a polygon around the objects in each frame
2. Model-Reconstruction – builds a force dynamic model of each polygon scene, determining grounding, attachment, and depth/layering
3. Event-Classification – determines over which intervals various primitive event types are true and from that data, over which intervals various compound event types are true

2.4 X-Schemas, F-Structs, and Model-Merging for Verb Learning

Bailey (1997) has developed a computational model of the role of motor control in verb acquisition. He argues that proprioception, which is knowledge of the body's own state, is linked to the acquisition of action verbs. In fact, he maintains that grounding action verbs in the motor-control system constrains the variety of lexical action categories and makes verb acquisition tractable. Bailey introduces the executing schema (x-schema) as a mechanism that can represent and carry out verbal commands, and feature structures (f-structs) as a mechanism for linking x-schema activities to related linguistic features.

X-schemas are formal representations of sequences of motor control actions. In Bailey's model, x-schemas are modeled as Petri nets with extensions to handle the passing of parameters.

In order to connect x-schemas to verbs, the linking feature structure (f-struct) is introduced. The f-struct is an intermediate set of features that allows a layer of abstraction between the individual motions of an action and the action verb that describes them. An f-struct is a list of feature-value pairs represented in a table with two rows. Each pair maps to a column with the feature located in the top row and the value in the bottom row. Bailey experientially determined a list of twelve features for his system comprised of eight motor control features and four perceived world state features.

Bailey's system performs verb acquisition using an algorithm that develops a lexicon of word senses based on a training set of verbs and linking f-structs summarizing that verb. Verb learning becomes an optimization problem to find the best possible lexicon given the training examples. Bailey terms this approach for merging word senses, *model-merging*, and implements a solution using a hill-climbing algorithm.

3 EBLA Model

The EBLA Model (Pangburn 2002) operates by observing a series of "experiences" in the form of short movies. Each movie contains a single event such as an arm/hand picking up a ball, and takes the form of either an animation or an actual video. The model detects any significant objects in each movie and determines what, if any, relationships exist among those objects. This information is then stored so that repeatedly occurring objects and relations can be identified across multiple experiences.

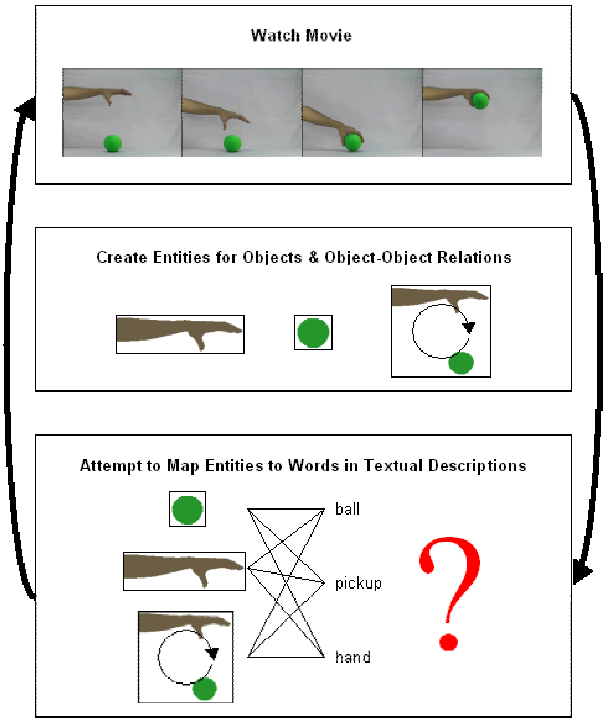

Figure 1. Method Used by EBLA to Process Experiences

As part of each experience, EBLA receives a textual description of the event taking place. These descriptions are comprised of protolanguage such as "hand pickup ball." To acquire this protolanguage, EBLA must correlate the lexical items in the descriptions to the

objects and relations in each movie. Figure 1 provides a graphical representation of the method used by EBLA to process experiences.

3.1 Model Abstractions and Constraints

The EBLA Model has been constrained in several ways. First, the model's perceptual capabilities are limited to a two-dimensional vision system that reduces objects to single color polygons.

Second, the model has not been provided with any audio processing capabilities. Because of this, all experience descriptions presented to or generated by EBLA are textual.

Third, the model only attempts to acquire a proto-language of nouns and verbs. Thus, syntax, word order, punctuation, etc. do not apply. This conforms with early human language acquisition since children do not begin to use phrases and clauses until somewhere between eighteen and thirty-six months of age (Calvin and Bickerton 2001).

The final constraint on EBLA is that it only operates in an unsupervised mode. This means that the model does not receive any sort of feedback regarding its accuracy. This is definitely a worst-case scenario since children receive frequent social mediation in all aspects of development.

3.2 Experiences Processed by the EBLA Model

The experiences processed by the EBLA Model are based on simple spatial-motion events, and take the form of either animations or real videos. Each experience contains an arm/hand performing some simple action on a variety of objects. For the animations, the actions include *pickup, putdown, touch,* and *slide*, and the objects include a green ball and a red cube (see figure 2). For the real videos, the actions include *push, pull, slide, touch, tipover, roll, pickup, putdown, drop,* and *tilt*, and the objects include several colored bowls, rings, and cups, a green ball, a dark blue box, a blue glass vase, a red book, and an orange stuffed Garfield cat (see figure 3).

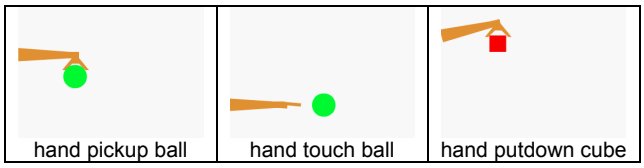

Figure 2. Frames from Various Animations Processed by EBLA

All of the videos were shot two to three times from both the left and right side of a makeshift stage. Angle of approach, grasp, and speed were varied at random. Multiple actions were performed on each object, but the actual object-event combinations varied somewhat based on what was feasible for each object. Dropping the glass vase, for example, seemed a bit risky.

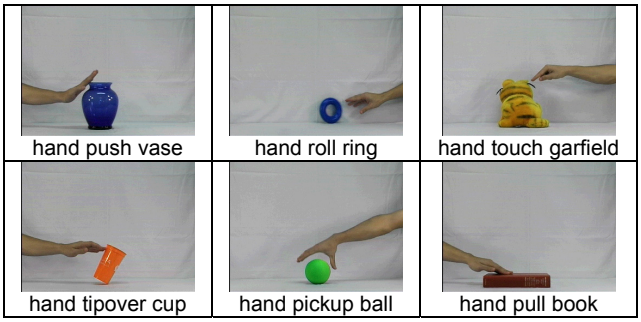

Figure 3. Frames from Various Videos Processed by EBLA

3.3 Entity Recognition

The EBLA Model has a basic perceptual system, which allows it to "see" the significant objects in each of its experiences. It identifies and places polygons around the objects in each video frame, using a variation of the mean shift analysis image segmentation algorithm (Comaniciu 2002). EBLA then calculates a set of static attribute values for each object and a set of dynamic attribute values for each object-object relation. The sets of attribute-value pairings are very similar to the linking feature structures (f-structs) used by Bailey (1997).

Each unique set of average attribute values defines an *entity*, and is compared to the entities from prior experiences. In order to match existing entities with those in the current experience, the existing entity must have average values for *all* attributes within a single standard deviation (σ) of the averages for the current entity. When this occurs, the current entity is merged with the existing entity, creating a more prototypical entity definition. Otherwise, a new entity definition is established.

To prevent entity definitions from becoming too narrowly defined, a minimum standard deviation (σ_{min}) is established as a percentage of each average attribute value. In essence, σ_{min} defines how much two entities must differ to be considered distinct, and thus can have a significant impact on the number of unique entities recognized by EBLA.

Both the object and relation attributes for EBLA were determined experimentally based on data available from the computer vision system. To aid in the debugging and evaluation of EBLA as well as to restrict any assumptions about early perception in children, an effort was made to keep the attributes as simple as possible. The five object attributes and seven relation attributes calculated by EBLA are briefly described in table 1.

Entity	Type	Description
area	object	area (in pixels) of a given object
grayscale value	object	grayscale color of object (0-255)
number of edges	object	number of edges on polygon tracing object
relative centroid (x)	object	horizontal coordinate of object's center of gravity relative to the width of a bounding rectangle around the object
relative centroid (y)	object	vertical coordinate of object's center of gravity relative to the height of a bounding rectangle around the object
contact	relation	Boolean value indicating if two objects are in contact with one another
x-relation	relation	indicates whether one object is to the left of, on top of, or to the right of another object
y-relation	relation	indicates whether one object is above, on top of, or below another object
delta-x	relation	indicates whether the horizontal distance between two objects is increasing, decreasing, or unchanged
delta-y	relation	indicates whether the vertical distance between two objects is increasing, decreasing, or unchanged
x-travel	relation	indicates direction of horizontal travel for both objects
y-travel	relation	indicates direction of vertical travel for both objects

Table 1. Entity Attributes Calculated by EBLA

Because average attribute values are used to define entities, temporal ordering is not explicitly stored in EBLA. Rather, the selected relation attributes implicitly indicate how objects interact over time. For example, EBLA is able to distinguish between *pickup* and *putdown* entities using the average "delta-y" attribute value—for *pickup*, the vertical distance between the two objects involved is decreasing over the experience and for *putdown*, the vertical distance is increasing.

Currently, object entities are defined using all of the object attributes, and relation entities are defined using all of the relation attributes. There is no mechanism to drop attributes that may not be relevant to a particular entity. For example, grayscale color value may not have anything to do with whether or not an object is a ball, but EBLA would likely create separate entities for a light-colored ball and a dark-colored ball.

A variation of the model-merging algorithm employed by Bailey (1997) could be applied to drop attributes unrelated to the *essence* of a particular entity. Because EBLA currently uses a limited number of attributes, dropping any would likely lead to overgeneralization of entities, but with more attributes, it could be a very useful mechanism. Such a mechanism would also improve EBLA's viewpoint invariance. For example, when detecting a *putdown* object-object relation, EBLA is not affected by small to moderate changes in angle, distance, or objects involved, but is affected by the horizontal orientation. Dropping the "x-relation" and "x-travel" attributes from the *putdown* entity would remedy this.

Work is underway to determine how to incorporate a 3D graphics engine into EBLA in order to build a more robust perceptual system. While this would obviously limit the realism, it would allow for the quick addition of attributes for size, volume, distance, texture, speed, acceleration, etc. Another option is to develop new attribute calculators for the current vision system such as those employed by Siskind (2000) to determine force dynamic properties.

3.4 Lexical Acquisition

Once EBLA has generated entities for the objects and object-object relations in each experience, its final task is to map those entities to the lexemes (words) in protolanguage descriptions of each experience. Protolanguage was chosen because it is the first type of language acquired by children. The particular variety of protolanguage used for the EBLA's experience descriptions has the following characteristics:

1. Word order is not important, although the descriptions provided to EBLA are generally in the format: s*ubject-manipulation-object* (e.g. "hand touch ball").
2. Verbs paired with particles are combined into a single word (e.g. "pick up" becomes "pickup").
3. Words are not case-sensitive (although there is an option in EBLA to change this).
4. Articles (e.g. "a," "an," "the") can be added to descriptions, but are generally uninterpretable by EBLA.

It should be noted that EBLA is not explicitly coded to ignore articles, but since they are referentially ambiguous when considered as individual, unordered lexemes, EBLA is unable to map them to entities. Adding articles to the protolanguage descriptions generally slows down EBLA's average acquisition speed.

In order to map the individual lexemes in the protolanguage descriptions to the entities in each experience, EBLA must overcome referential ambiguity. This is because EBLA operates in a bottom-up fashion and is not primed with any information about specific entities or lexemes. If the first experience encountered by EBLA is a hand sliding a box with the description "hand slide box," it has no idea whether the lexeme "hand" refers to the *hand* object entity, the *box* object entity, or the *slide* relation entity. This same referential ambiguity exists for the "slide" and "box" lexemes. EBLA can only overcome this ambiguity by comparing and contrasting the current experience with future experiences. This process of resolving entity-lexeme mappings is a variation of the cross-situational learning employed by Siskind (1992; 1997).

For each experience, two lists are created to hold all of the unresolved entities and lexemes. EBLA attempts to establish the correct mappings for these lists in three stages:

1. Lookup any known resolutions from prior experiences.
2. Resolve any single remaining entity-lexeme pairings.
3. Apply cross-situational learning, comparing unresolved entities and lexemes across all prior experiences, repeating stage two after each new resolution.

To perform the first stage of lexical resolution, EBLA reviews known entity-lexeme mappings from prior experiences. If any match both an entity and lexeme in the current experience, those pairings are removed from the unresolved entity and lexeme lists.

The second stage operates on a simple process of elimination principal. If at any point during the resolution process both the unresolved entity and lexeme lists contain only a single entry, it is assumed that those entries map to one another. In addition, prior experiences are searched for the same entity-lexeme pairing and resolved if found. Since resolving mappings in prior experiences can generate additional instances of single unmapped pairings, the entire second stage is repeated until no new resolutions are made.

The third and final stage of resolution is by far the most complex and involves a type of cross-situational inference. Basically, by comparing the unresolved entities and lexemes across all experiences in a pair wise fashion, EBLA can infer new mappings. If the cardinality of the intersection or difference between the unmapped entities and lexemes for a pair of experiences is one, then that intersection or difference defines a mapping. In more formal terms:

1. Let i and j be any two experiences, $i \neq j$.
2. Let E_i and $E_j \in$ unmapped entities for i and j respectively.
3. Let L_i and $L_j \in$ unmapped lexemes for i and j respectively.
4. If $|\{E_i \cap E_j\}| = 1$ and $|\{L_i \cap L_j\}| = 1$ then $\{E_i \cap E_j\}$ maps to $\{L_i \cap L_j\}$.
5. If $|\{E_i \setminus E_j\}| = 1$ and $|\{L_i \setminus L_j\}| = 1$ then $\{E_i \setminus E_j\}$ maps to $\{L_i \setminus L_j\}$.
6. If $|\{E_j \setminus E_i\}| = 1$ and $|\{L_j \setminus L_i\}| = 1$ then $\{E_j \setminus E_i\}$ maps to $\{L_j \setminus L_i\}$.

To demonstrate how all three stages work together, consider the following example. If the model was exposed to an experience of a hand picking up a ball with the description "hand pickup ball" followed by an experience of a hand picking up a box with the description "hand pickup box," it could take the set differences discussed in stage three for the two experiences to resolve the "ball" lexeme to the *ball* entity and the "box" lexeme to the *box* entity. Assuming that these were the only two experiences presented to the model, it would not be able to resolve "hand" or "pickup" to the corresponding entities because of referential ambiguity. If the model was then exposed to a third experience of a hand putting down a ball with the description "hand putdown ball," it could resolve all of the remaining mappings for all three experiences. Using the technique discussed in stage one, it could resolve "ball" based on known mappings from the prior experiences. It could then take the set intersection with the unmapped items in either of the first two experiences to resolve "hand." This would leave a single unmapped pairing in each of the three experiences, which could be resolved using the process of elimination discussed in stage two. Note that taking the set difference rather than the intersection between the third and first or second experiences would have worked equally well to resolve "hand pickup" and "hand putdown."

4 Evaluation

EBLA was evaluated using three criteria. First, overall success was measured by comparing the number of correct entity-lexeme mappings to the total number of entities detected. Second, acquisition speed was measured by comparing the average number of experiences needed to resolve a word in comparison to the total number of experiences processed. Third, descriptive accuracy was measured by presenting EBLA with new, unlabeled experiences, and determining its ability to generate protolanguage descriptions based on prior experiences.

The test sets for EBLA were comprised of eight simple animations created using Macromedia Flash, and 319 short digital videos. While the results for the animations were somewhat better than those for the videos, only the results for the larger and more complex video test set will be presented here.

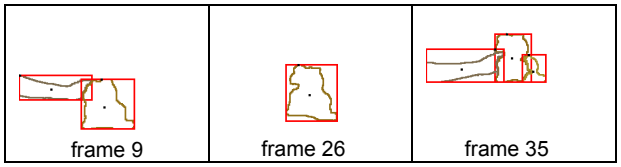

Figure 4. Polygon Traces from a Single Video Demonstrating Normal Segmentation, Undersegmentation, and Oversegmentation

Of the 319 videos, 226 were delivered to EBLA for evaluating lexical acquisition accuracy and speed and 167 were delivered to EBLA for evaluating descriptive accuracy. Videos were removed from the full set of 319 because of problems with over and undersegmentation in the vision processing system. Figure 4 demonstrates the types of problems encountered by EBLA's vision

system. It shows the polygon tracings for three frames from a single video shot with the Garfield toy. The frame on the left was correctly segmented, the frame in the middle was undersegmented where the hand has been merged into the background and essentially disappeared, and the frame on the right was oversegmented where the Garfield toy has been split into two objects.

To measure acquisition speed and accuracy, the 226 videos were delivered to EBLA at random, ten times for each of nineteen different minimum standard deviation (σ_{min}) values. The value of σ_{min} used to match the attribute values to existing entities was varied from 5% to 95% in increments of 5%.

Figure 5 shows the success rates for lexeme mappings for each of the nineteen σ_{min} values. For σ_{min} values of 5% and 10%, the acquisition success was only 76% and 85% respectively. This can be attributed to the amount of variation in the entities for the videos. A stricter matching criteria results in more unmatched entities. For all of the other σ_{min} values the acquisition success rate was better than 90% and as high as 95.8% for a σ_{min} value of 45%.

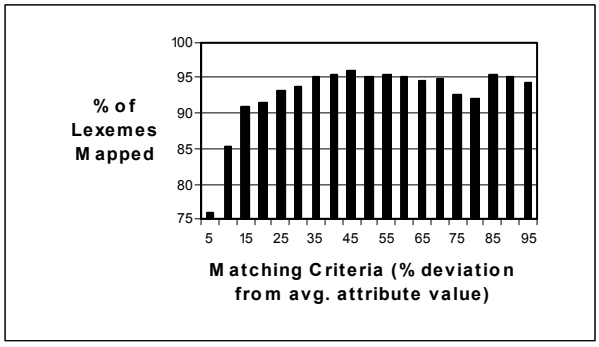

Figure 5. Lexeme Mapping Success Rates for Different Minimum Standard Deviations

Figure 6 displays the average acquisition speed for the videos. It indicates that for the first few videos, it took an average of over twenty experiences to resolve all of the entity-lexeme mappings. After about seventy-five experiences had been processed, this average dropped to about five experiences, and after about 150 experiences, the average fell below one.

To evaluate the descriptive accuracy of EBLA, 157 of the 167 best videos were randomly processed in acquisition mode and the remaining ten were processed in description mode. This scenario was run ten times for each of the same nineteen σ_{min} values used to evaluate acquisition success. The results are shown in table 2. It is important to note that for a given σ_{min} value, EBLA often returned multiple "matching" lexemes. When this happened, both the correct and incorrect lexemes were scored pro-rata.

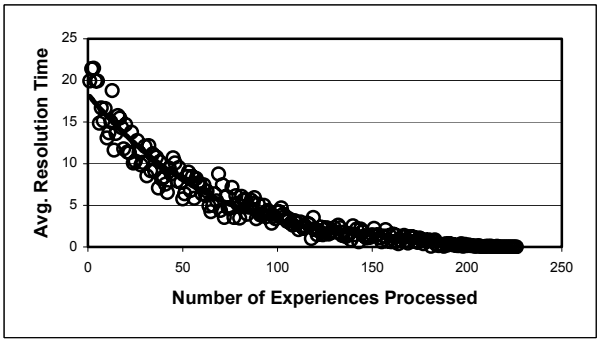

Figure 6. Average Lexical Acquisition Time for Videos

σ_{min}	% Correct	% Incorrect	% Unknown
5	50.33	9.00	40.67
10	57.22	14.11	28.67
15	65.33	16.00	18.67
20	56.07	25.27	18.67
25	57.44	27.89	14.67
30	62.94	27.73	9.33
35	59.30	35.03	5.67
40	63.14	30.52	6.33
45	60.95	34.05	5.00
50	50.83	41.17	8.00
55	55.04	40.62	4.33
60	48.39	45.94	5.67
65	46.21	49.46	4.33
70	49.96	45.38	4.67
75	43.63	53.03	3.33
80	44.42	50.91	4.67
85	46.45	50.55	3.00
90	45.04	52.62	2.33
95	39.51	54.49	6.00

Table 2. Accuracy of Video Descriptions

For the lower values of σ_{min}, there were very few incorrect descriptions, but many entities did not map to a known lexeme. As σ_{min} was increased, the situation reversed with almost every entity mapping to some lexeme, but many to the wrong lexeme. The most accurate descriptions were produced for a σ_{min} value of 15% where just over 65% of the entities were described correctly. These are reasonably good results considering the amount that any given entity varied from video to video, especially the object-object relation entities. For a full discussion of both the animation and video results for EBLA see chapter 6 of Pangburn (2002).

5 Conclusion

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. In addition, because EBLA operates in an online fashion, it does not require an explicit training phase.

EBLA performed very well on the entity-lexeme mapping task for both the animations and the videos, achieving success rates as high as 100% and 95.8% respectively. EBLA was also able to generate descriptions for the animations and videos with average accuracies as high as 96.7% and 65.3%. The 65.3% is still quite good when compared to the approximately 15% average success rate obtained by generating three word descriptions at random from the pool of nineteen lexemes processed by EBLA.

While the initial results from the EBLA system are encouraging, much development and evaluation remains to be done. Adding new attribute calculators along with a mechanism for dropping extraneous attributes would likely make EBLA's entity definitions more robust and facilitate the acquisition of additional nouns and verbs as well as other parts of speech. Since there is nothing in the design of EBLA that prevents it from processing videos with more than three entities/lexemes, it should be thoroughly tested using more complex experiences and/or descriptions

As mentioned in the introduction, one of the primary goals of EBLA has been to develop an open system that would be relatively easy for others to use and extend. To that end, EBLA was written entirely in Java with a PostgreSQL relational database for storage of all experience parameters, intermediate results, attribute definitions and values, lexemes, entity definitions, and entity-lexeme mappings. EBLA has been released on SourceForge at http://sourceforge.net/projects/ebla/. For more information on EBLA, visit http://www.greatmindsworking.com

References

Brian E. Pangburn. 2002. Experience-Based Language Acquisition: A Computational Model of Human Language Acquisition, Ph.D. thesis, Louisiana State University, LA.

David R. Bailey. 1997. When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs, Ph.D. thesis, University of California, Berkeley, CA.

Deb Kumar Roy. 1999. Learning Words from Sights and Sounds: A Computational Model. Ph.D. thesis, Massachusetts Institute of Technology.

Deb Kumar Roy. 2000. Learning Visually Grounded Words and Syntax of Natural Spoken Language. In *Evolution of Communication Journal* 4, no. 1 (April): 33-57.

Dorin Comaniciu. 2002. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, no. 5 (May): 603-619.

Emily Smith. 1999. The Performance of Prekindergarten Children on Representational Tasks Across Levels of Displacement. Ph.D. thesis, Louisiana State University.

George Lakoff. 1990. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: The University of Chicago Press.

Janet A. Norris and Paul R. Hoffman. 2002. Language Development and Late Talkers: A Connectionist Perspective. In *Connectionist Approaches to Clinical Problems in Speech and Language: Therapeutic and Scientific Applications*, ed. Raymond G. Daniloff, 1-109. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jeffrey M. Siskind. 1992. Naïve Physics, Event Perception, Lexical Semantics, and Language Acquisition, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Jeffrey M. Siskind. 1997. A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings. In *Computational Approaches to Language Acquisition*, ed. Michael Brent, 39-91. Amsterdam, Netherlands: Elsevier Science Publishers.

Jeffrey M. Siskind. 2000. Visual Event Classification via Force Dynamics. *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA.

Jeffrey Mark Siskind and Quaid Morris. 1996. A Maximum-Likelihood Approach to Visual Event Classification. In *Proceedings of the Fourth European Conference on Computer Vision (ECCV '96)* Vol. 2, 347-360. New York, NY: Springer-Verlag.

John L. Locke. 1993. *The Child's Path to Spoken Language*. Cambridge, MA: Harvard University Press.

Katherine Nelson. 1998. *Language in Cognitive Development: The Emergence of the Mediated Mind*. Cambridge, UK: Cambridge University Press.

Luc Steels and Frederic Kaplan. 2000. AIBO's First Words: The Social Learning of Language and Meaning. In *Evolution of Communication Journal* 4, no. 1 (April): 3-32.

Steven Pinker. 2000. *The Language Instinct: How the Mind Creates Language*. New York, NY: William Morrow and Company.

Webster's Ninth New Collegiate Dictionary. 1989. s.v. "apple."

William H. Calvin and Derek Bickerton. 2001. *Lingua ex Machina: Reconciling Darwin and Chomsky with the Human Brain*. Cambridge, MA: MIT Press.