Grounding Color and Shape Words in Visual Context

Celia Abernathy and Lisa Bao

May 11, 2012

Abstract

In this paper we present an adaptive learning system which uses a three-phase recurrent neural network, based on [8], to learn simple color and shape vocabulary. The network is given simulated images and complete English sentences as input; it is expected to output images corresponding to a given sentence and sentences corresponding to a given image. By grounding the vocabulary in a sensory context, our system is able to infer both semantic meaning and syntactic organization. The network performed very well on both production and comprehension of language, although the benefits of a multi-phase structure remain unclear.

1 Symbol Grounding

Adaptive robotics and artificial intelligence systems have come a long way since the heady, optimistic days of Classical AI, but the computer science community remains some time away from creating a truly intelligent, autonomous robot. One approach to the AI problem, inspired by developmental psychology and cognitive science, focuses on bootstrapping adaptive systems which can learn implicitly about the real world and gradually develop more complex, human-like capabilities. This approach, called connectivism, supports a cognitive theory of the mind based on existing brain structures in nature [6].

Within this field of research, natural language comprehension and production is another intractable problem. Human babies are able to learn any natural language to native fluency, but creating an autonomous system to do the same is currently impossible. The issue of symbolic grounding arises when systems attempt to learn natural language: how can a robot understand what words (symbols) mean in relation to the real world?

As a solution to the symbol grounding problem, Harnad proposed a hybrid symbolic/connectivist model which grounds basic symbols directly into sensorimotor input and categorizes them to provide the scaffolding for transfer of the grounded representation to higher-order symbols [6]. Moreover, he asserted that the only viable method of merging symbolic and connectivist theory is bottom-up, beginning first with the intrinsic meaning of the real-world representation.

1.1 Neural Networks

Our system is implemented as an artificial neural network with a standard backpropagation algorithm. The structure of neural networks is inspired by networks of interconnected neurons observed in the brain, "where each unit takes a number of real-valued inputs (possibly the output of other units) and produces a single real-valued output (which may become the input to many other units)" [7, p.82]. Neural networks, especially in conjunction with the backpropagation algorithm,

are commonly used in machine learning. A network consists of interconnected nodes. Backpropagation learns the network weights by propagating inputs forward through the network (called "feed-forward") and propagating error backward; the weights of connections between nodes are adjusted based on this error. The learning algorithm is guaranteed to converge only to a local minimum, but remains highly effective despite this constraint.

The adaptive system we propose below is inspired by Alex, an African gray parrot who worked with comparative psychologist Irene Pepperberg for three decades to learn about 150 English words [3]. Alex could categorize his vocabulary, demonstrate understanding of colors and shapes, and count to small numbers. Our neural network, christened Baby Alex, does not attempt to replicate all of Alex"s achievements; rather, it simplifies the task to comprehending and producing English sentences which describe a limited range of images. Rather than being given explicit instruction about color and shape, Baby Alex must use syntactic information to connect the meanings of words to the different qualities of an image.

1.2 Related Work

Our system derives heavily from Plunkett, et al.'s work on symbol grounding, where a neural network takes images and labels as input [8]. Their three-phase auto-associative neural network trains first on only the image branch, then only the label branch, and finally the entire network. Each modality's hidden units, along with the three-phase training system, allows the network to connect internal representations of labels and images. Including the combined third phase allows the network to complete tasks which test both its comprehension of images and its production of labels. The images are represented as random dot patterns and "seen" in two dimensions using a retinal pre-processor; each image is paired with a label represented as an arbitrary orthogonal vector. The network is also tested on label understanding by presenting it with either a blank image or a blank label and checking its image or label output.

Plunkett, et al. based their work in part on observations from childhood language development. For example, children develop language in stages, first appearing to reach a plateau before suddenly experiencing a vocabulary growth spurt. Words become decontextualized, and their grounding moves from specific contexts to a conceptual representation. Indeed, Plunkett, et al. found in their results that a comprehension growth spurt occurred early on in the network's training. Also parallel to human development, an asymmetry of accuracy was observed between comprehension and production.

Around the same time, Chauvin was working on a similar connectionist model of symbol emergence [4]. He used an auto-associative network to study categorization during the semantic acquisition period of language development, when "an entity in a modality becomes consistently mapped to another entity in a different modality" (580). The PDP network is given two modalities of input, auditory and visual, within categories each containing a set of images and a label. The images are random dot patterns pre-processed to form a two-dimensional activation pattern and the category labels are abstracted to simply A, B, C, or D; each is presented both alternately and in unison to the network during its learning period. Like Plunkett, et al., Chauvin justified his results with theories of first-word acquisition in children.

More recently, Cangelosi, et al. in 2007 used a three-stage training process to directly ground language with a robot's action representations [2]. Each symbol is represented as a perceptual, sensorimotor, social, or other internal categorization, and different categories of symbols maintain logical syntactic relationships with other symbols. Using these syntactic relationships, the robot is

able to transfer the meaning of lower-order grounded symbols onto higher-order symbols, a process called symbol grounding transfer. Cangelosi, et al. based their work on the cognitive science principle that "robots acquire words through direct interaction with their physical and social world, so that linguistic symbols do not exist as arbitrary representations [but as] intrinsically connected to behavorial or cognitive abilities" (66).

Our system, Baby Alex, extends Plunkett, et al.'s approach by adding recurrence to the network architecture. Based on Elman's 1990 paper, we add a context layer that contains a copy of the previous step"s activations and feeds into the hidden layer. Instead of showing the network discrete words, as in Plunkett, et al., we provide full sentences, one word at a time. The context layer allows Baby Alex to recall the previous word it has seen, thus giving it a sense of syntactic structure.

2 Implementation

Baby Alex is founded upon a feed-forward neural network. The network trains on examples, and a standard backpropogation algorithm is used to adjust the connection weights. Most of the neural network functions are handled by the conx module of Pyro robotics [1]. However, we extended the code in two ways: adding recurrence and adding a three-phase training structure.

2.1 Network Structure

Figure 1 shows the underlying structure of Baby Alex, which replicates the structure of Plunkett, et al., albeit with a differing number of units. The network simultaneously processes two inputs: an image and a word. Each modality has its own hidden units that allow the network to create a representation for the input before sending that representation to the shared hidden layer. Each

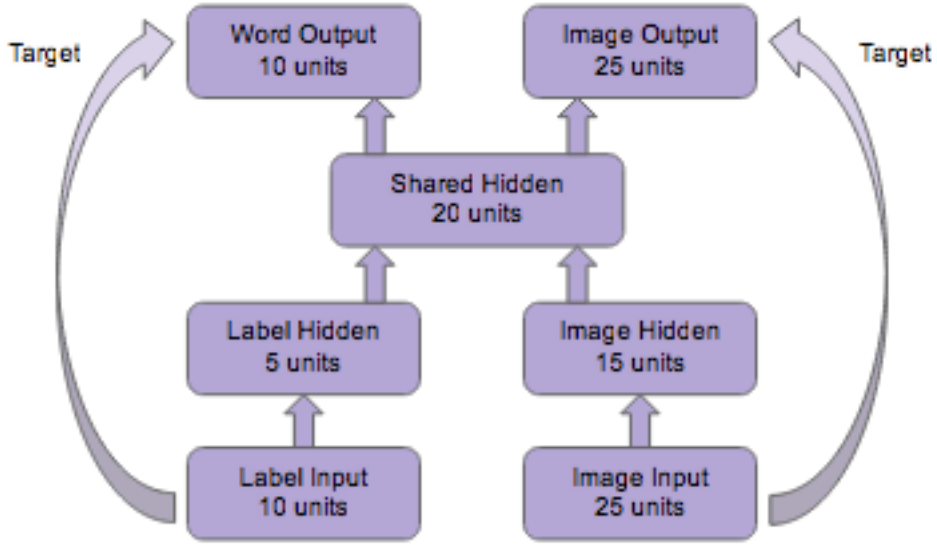


Figure 1: The two-sided, auto-associative structure, similar to that described in [8], which forms the basis of Baby Alex.

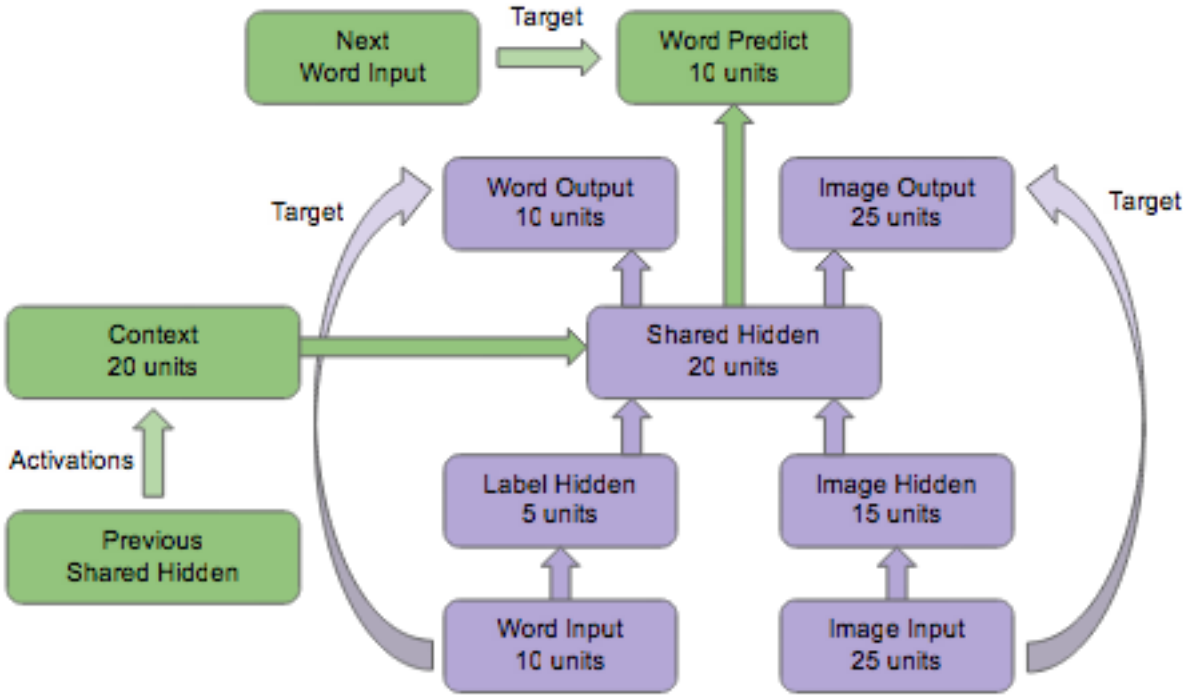


Figure 2: Baby Alex's network architecture, including a recurrent context unit and an output whose task is prediction.

modality likewise has an output layer whose task is to reproduce the input. This task, called auto-association, is not trivial; because the hidden layers are smaller than the input and output layers, the network must organize input into simpler representations in order to successfully reproduce it on the output nodes [8].

Our system combines this architecture with a recurrent structure to give Baby Alex an understanding of sentence syntax. We modified the architecture by adding a context layer that feeds into the shared hidden layer. At every step, this context layer is given the activations of the shared hidden layer from the previous step. This allows the network to remember what happened in the immediate past, as described by Elman [5].

In order to test Baby Alex's ability to produce appropriate words in context, we added an additional output layer. The target for this layer was updated on every time step to match the next word input. Hence, Baby Alex was trained to predict the next word. Figure 2 shows the final network architecture.

2.2 Task

As mentioned in Section 2.1, Baby Alex accepts both images and words as inputs. For the sake of simplicity, each image is a string of 25 numerical inputs that represents an object. Objects differ in color and shape: red, green, or blue is represented by different float values, while a square, line, triangle, or circle is represented by the arrangement of colored units and white (0.0) units on a 5 by 5 grid, which is then flattened. This schematic can be seen in Figure 3. By combining color and

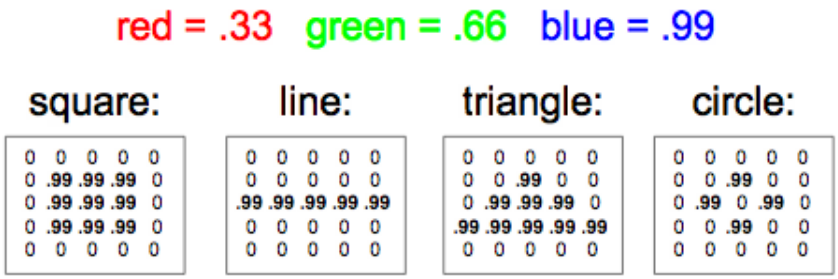


Figure 3: An illustration of the three possible colors and four possible shapes that combine to form image representations.

shape attributes in all possible ways, 12 distinct objects are created.

Word inputs are translated into a set of orthogonal vectors in which all units are 0.0 except for one unit set to 1. These words make up five-word sentences that describe each object, such as "This is a blue square" or "This is a red triangle." Baby Alex is shown only one word at a time. Each sentence is associated with the matching image, and that image appears alongside every word in the sentence. The auto-association task of replicating the inputs should be straightforward, but the word prediction task is more complex. Baby Alex needs to both understand the structure of a sentence and be able to translate visual data into words.

2.3 Training

Plunkett, et al. described a three-phase training structure which gave better results than straightforward neural network training [8]. In each phase, only certain layers of the network would be active – only those layers would propagate activation forward, and only weights between those layers would be changed by backpropagation. In the first phase, only layers concerned with the label are active; in the second phase, the layers concerned with the image are active; and in the third phase, all layers are active. Plunkett, et al. suggest that the three-phase training structure acts as an attention-switching process, allowing the network to attend to the two modalities separately before drawing associations between them. We implemented similar phases for Baby Alex, as shown in figure 4.

In Plunkett, et al.'s implementation, the network cycled through all of these phases for each word/image pair. However, several different phase levels are possible. We tested various levels at which to cycle through the phases:

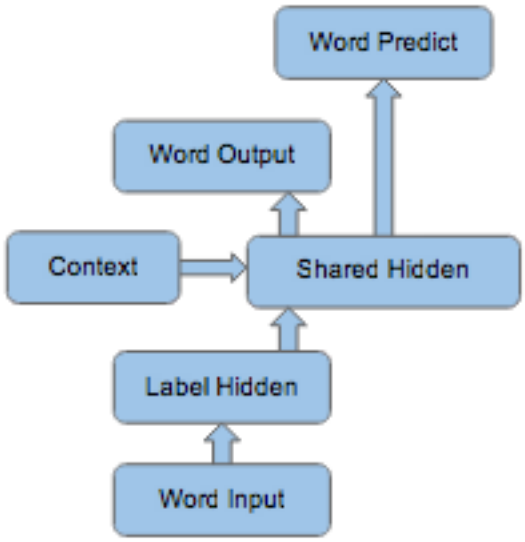
Step: Cycle through phases on each word/image pair

Sentence: Cycle through phases on each sentence

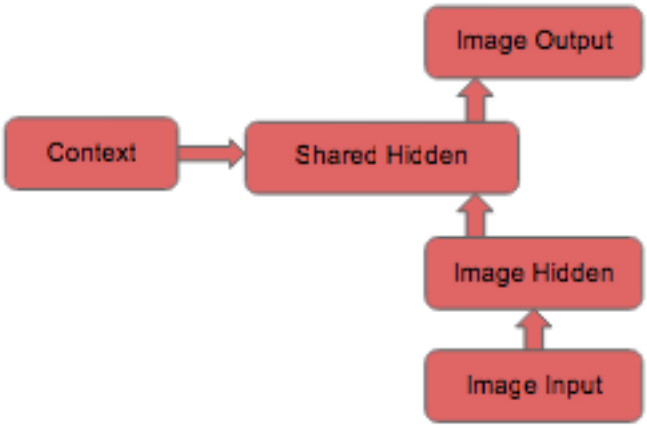
Sweep: Cycle through phases on each sweep through every sentence in the training data

Phaseless: No phases; all layers are always active

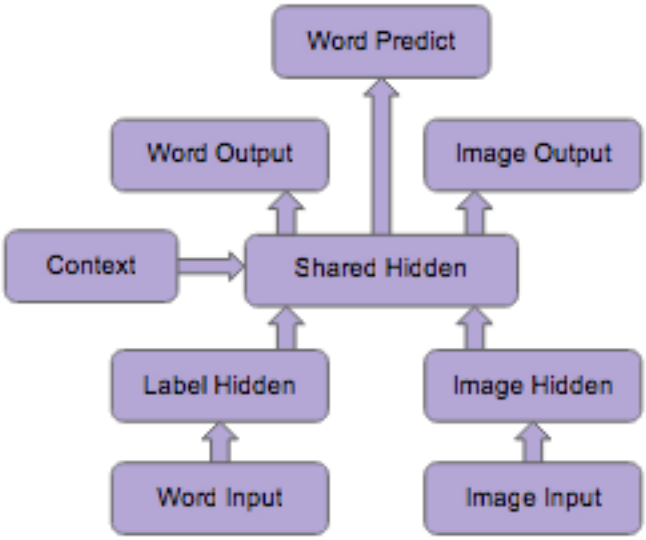
Each of the 12 unique sentences was 5 words long and attached to an image. Baby Alex trained on noisy versions of these images, so it never saw the prototypical objects during training. The noisy versions were created by adding a random number between -.05 and .05 to each image input



(a) Phase 1



(b) Phase 2



(c) Phase 3

Figure 4: Active layers during the each phase. Only these layers feed activation forward or have their weights changed by backpropagation.

unit, which was done 5 times per sentence/image pair, resulting in a total of 60 sentences in the training data. The order of the sentences was chosen at random.

By training Baby Alex to predict the next word, we implicitly taught it to produce descriptive sentences, as it must come up with all of the words in the sentence. In order to reward sentence comprehension, the training data also included several blank image sentences. For these sentences, the words are presented alongside a blank image (an input of all 0's), but the target for the image output is the representation matching the verbal description. We added 5 versions of each blank image sentence to the training data for a total of 120 sentences.

Baby Alex was trained for 100 sweeps, where a sweep is one run through each sentence in the training data. After every 5 sweeps, we stopped and tested Baby Alex according to the experimental procedures described in Section 3.1. We repeated this process 10 times for each of the 4 phase level conditions.

3 Experiment

Two major parts of understanding and using language are syntax and semantics. Through recurrence, Baby Alex should be able to understand and reproduce the structure of simple descriptive sentences. Likewise, through a three-phase structure, Baby Alex should be able to understand the meaning of color and shape words by grounding them in its perception of images. Therefore, we hypothesized that Baby Alex will be able both to produce a sentence describing an image and to comprehend a descriptive sentence by imagining the object described. We also expected that training the network using a three-phase method would give better results than using a phaseless method.

3.1 Evaluation

Because our goals included both production and comprehension, we tested Baby Alex on two special sets of testing data. Each set had only 12 sentences, one for each prototypical image. We averaged Baby Alex's accuracy rate on the 12 sentences to arrive at a final result for that test. The production test used normal image/sentence combinations and the comprehension test used a blank image versions of the sentences.

First, production was tested by measuring the accuracy of the word prediction output. We calculated the percentage correct based on the difference between each unit of the output and the target. To see how Baby Alex's performance varied throughout a sentence, we separated the results for each word. We were especially interested in how it performed on the color and shape words. The results for the words this, is, and a were very similar to each other, so we averaged them together in our collated results.

Second, comprehension was tested in two parts: color comprehension and shape comprehension. To measure color comprehension, we examined only two of the output units: the two units that are "on" (that is, colored instead of white) for every shape. This way, if Baby Alex got the shape wrong, the color would still be measured correctly. The percentage correct for these two units was calculated similarly to the label units. To measure shape comprehension, we checked which units were on (having an activation of ≥ 0.1) versus off. The percentage correct was simply a quotient: the number of units correctly on or off, divided by the total number of units. Comprehension was only measured on the last word of the sentence, at which point Baby Alex had all the necessary

	Production			Comprehension	
Type	Color	Shape	Others	Color	Shape
Step	98.764	99.326	99.810	86.817	93.100
Sentence	98.100	98.330	99.687	87.278	94.733
Sweep	97.779	98.898	99.715	91.715	97.033
Phaseless	98.769	99.164	99.738	93.889	96.967

Table 1: Average results over 10 trials for production (color words, shape words, and an average of the other words: "this", "is", and "a") and comprehension

information to imagine the image.

3.2 Results

Table 1 shows the average percentage correct for both production and comprehension. Baby Alex was successful both at predicting the next word when looking at an image and at producing an image whose color and shape are close to what a sentence describes.

However, networks that cycled through phases at different levels had somewhat different results. Figure 5 averages the results in table 1. The step network was marginally better than the phaseless network at production, while the other networks were slightly worse at production. The results concerning the step and phaseless network are similar to the results of Plunkett, et al., who found

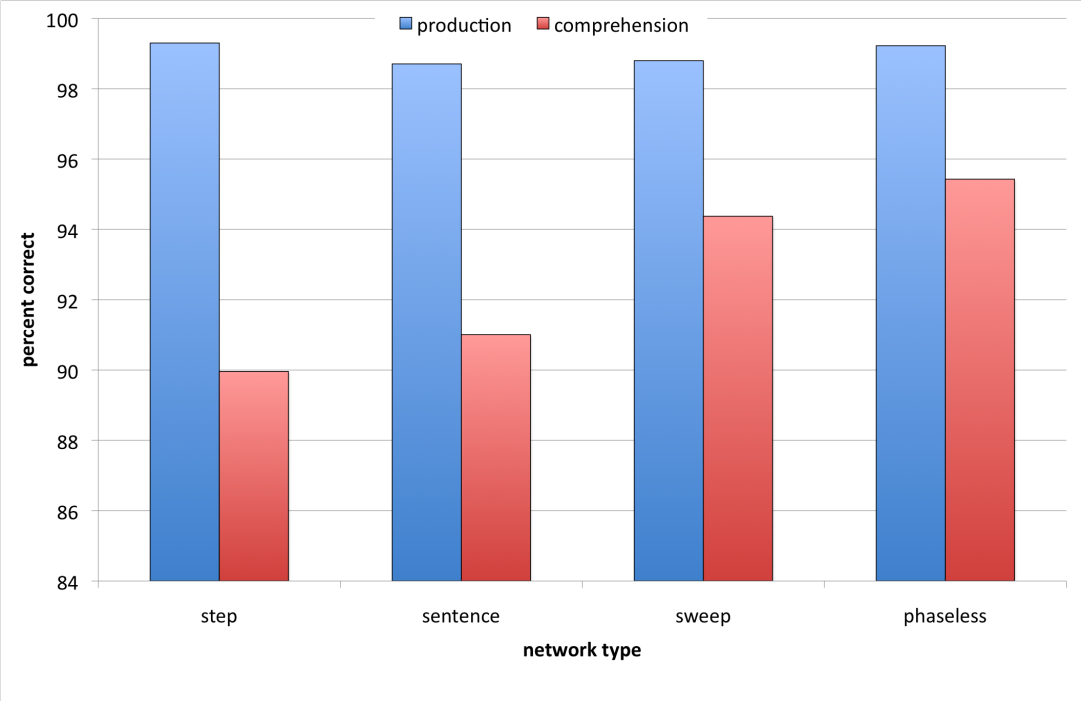


Figure 5: Average results over 10 trials for production and comprehension, for each phase level type.

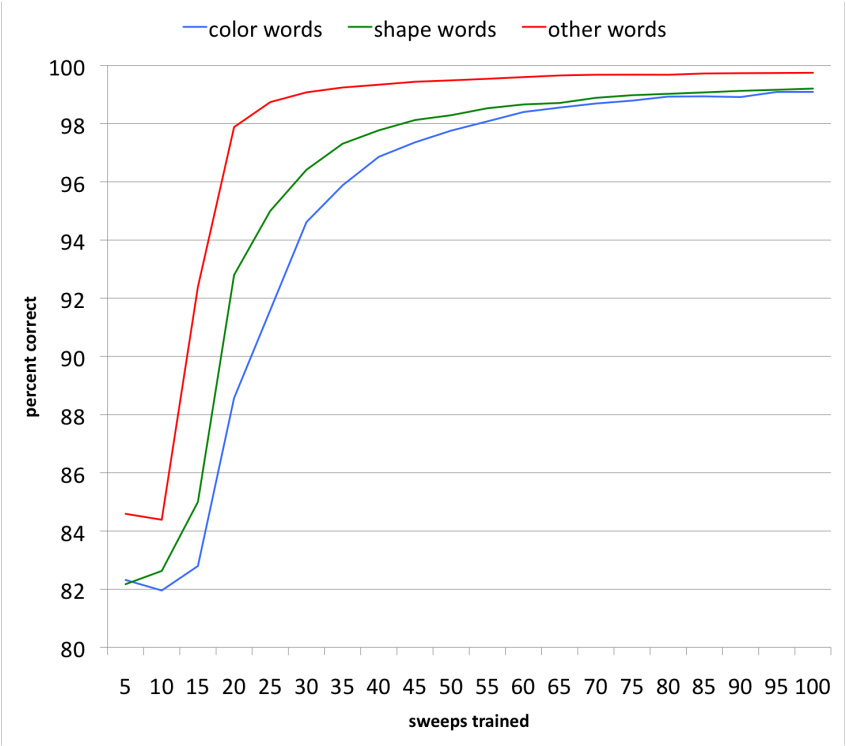


Figure 6: Production accuracy for a phaseless network.

Figure 7: Comprehension accuracy for a phaseless network.

that the three-phase training regime (which cycled at the step level) was marginally better than a phaseless training regime [8].

Although the step network achieved the highest percentage correct in production, it was the worst at comprehension. The phaseless network was the most successful at comprehension. In fact, Figure 5 shows a trend in comprehension results: the longer a network trains uninterrupted (that is, the longer each phase), the better it does on comprehension. More investigation is needed to investigate this pattern and its cause.

Clearly, cycling through phases during training does not produce a significant overall advantage for the network. For comprehension, phases actually reduce Baby Alex's effectiveness; for production, only the step network shows an improvement over the phaseless network, and the increase is marginal. Furthermore, there does not appear to be an obvious relationship between high performance on production and high performance on comprehension among the different phase level conditions.

Since the results about the different phase levels are so inconclusive, we will use the phaseless network as an example to examine the production and comprehension results in further detail. Figure 6 shows how the network improved as it trained throughout a representative trial. For instance, the words this, is, and a were learned quickly and accurately. Moreover, shape words were easier for Baby Alex to grasp than color words.

Figure 7 shows the comprehension results for a trial of the phaseless network. These results are more variable than the production results, which is unsurprising given that the task involved producing an output 25 units in length with finer differentiation (between .33 and .66, for instance), versus a 10-unit-long vector comprised of 0's and 1's. Like for production, the comprehension of shape was an easier task than color.

4 Discussion and Conclusions

Our original goal for Baby Alex was to combine Plunkett, et al.'s system for grounding semantic meaning with Elman's system for understanding syntax to produce and comprehend descriptive sentences. Specifically, we wished to teach Baby Alex how to use autonomously-learned syntactic information (for example, the fact that color words always come before shape words in a simple descriptive sentence) to associate meaning with the physical features of an image.

Baby Alex was successful at this task: it became proficient at both production and comprehension. Even though it had never seen the prototypical objects before, it could describe them accurately, and it could imagine objects correctly. However, the three-phase training method that we hoped would improve the network's learning failed to provide any clear benefit when compared to a normal phaseless training system. This experiment was not enough to fully understand the impact of the three-phase structure.

These results are significant as a prototype of a system to adaptively learn language. We have shown that a recurrent context layer "memory" can successfully extend the behavioral abilities of a system which also grounds abstract language in sensory input. Both kinds of abilities will be crucial to the design of an autonomous, intelligent robotic system.

However, this experiment was limited in scope. The images and sentences were both simplified. In future work, it would be interesting to expand Baby Alex's vocabulary and grammar, introducing new descriptors or new sentence structures. Ultimately, this system should be tested on real images rather than simplified representations, as a step towards the final goal of learning and using language

in the real world. Another avenue is to investigate the different phase cycles, in order to understand their impact on the network and how they might improve performance.

Acknowledgments

We sincerely thank our professor, Lisa Meeden, for her continued guidance and advice throughout this project.

References

[1] Douglas Blank, Deepak Kumar, Lisa Meeden, and Holly Yanco. Pyro: A python-based versatile programming environment for teaching robotics. *Journal of Educational Resources in Computing*, 2004.

[2] Angelo Cangelosi, Vadim Tikhanoff, Jos Fernando Fontanari, and Emmanouil Hourdakis. Integrating language and cognition: A cognitive robotics approach. *IEEE Computational Intelligence Magazine*, 2(3):65–70, 2007.

[3] Benedict Carey. Alex, a parrot who had a way with words, dies. *New York Times*, 42:335–346, September 2007.

[4] Yves Chauvin. Toward a connectionist model of symbolic emergence. In *Program of the Annual Conference of the Cognitive Science Society*, pages 580–587, 1989.

[5] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[6] Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[7] Tom Mitchell. Artifical neural networks. In *Machine Learning*, pages 81–127. McGraw Hill, 1997.

[8] Kim Plunkett, Chris Sinha, Martin F. Moller, and Ole Strandsby. Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4(3-4):293–312, 1992.