5th International Conference on Development and Learning
and on Epigenetic Robotics
Aug 13-16, 2015. Providence, RI USA

Using spatial representations in gesture to facilitate early word learning: A priming process model

J. Gregory Trafton
Naval Research Laboratory
Washington, DC
greg.trafton@nrl.navy.mil

Anthony M. Harrison
Naval Research Laboratory
Washington, DC
anthony.harrison @nrl.navy.mil

Wallace E. Lawson
Naval Research Laboratory
Washington, DC
ed.lawson@nrl.navy.mil

***Abstract*—As children learn to speak, they also gesture; previous empirical work has suggested that there is a direct link between the two. In this paper, we propose a priming process model that uses gesture to facilitate language. Our model uses the ACT-R/E cognitive architecture and uses a combination of repetition naming and priming from gesture spatial representations to increase the probability that a word will be remembered. Our model simulates 11 months of learning and runs on an embodied platform.**

***Index Terms*—ACT-R/E, cognitive robotics, gesture learning, human robot interaction, priming.**

I. Introduction

Language and gesture are intimately tied together [1-3]. Children typically point before they speak [4] and gesture becomes more elaborate as children grow older. Most researchers believe that gesture and language are separate systems, but that they co-develop; it is the nature of exactly how they co-develop at a very young age that we focus on in this report.

Gesture is also a current research topic in robotics, especially human-robot interaction (HRI) and developmental robotics. Most of the work in HRI focuses on creating robots that gesture naturally so that a person can understand them [5] or creating joint attention with gesture as a possible modality [6; 7].

In the last decade, evidence has mounted that children not only use gesture before language, but that gesture actually facilitates language development [8]. In fact, researchers have recently built embodied models that use gesture to facilitate language learning [9; 10]

Sheldon and Lee (2010, 2011) developed a robotic system that was based on Iverson and Goldin-Meadow's (2005) work. Their system starts with motor babbling and maps motor actions to visual objects in the environment and then progresses to pointing and one and two word speech. Their system is based on computational formalisms of Piagetian schemas and is able to learn series of precondition/postcondition pairs through schema chaining. Sheldon and Lee's system is able to learn with the help of an active participant how to pick up, reach, point, and label different objects in the environment over time. Their system does not, however, attempt to accurately model human development.

It is a mantra in the modeling community that no model is perfect; future models attempt to improve upon past models. The Sheldon and Lee model is excellent, but its biggest weakness in our opinion is that it does not maintain cognitive plausibility or constraints. In order to show cognitive plausibility, we (1) use and integrate a variety of cognitively plausible mechanisms (e.g., models of human memory, spatial representation, etc.), (2) run models using a similar experimental paradigm, and (3) match experimental data using those mechanisms within the constraints of the experimental paradigm.

The data we attempt to match is an experiment by Iverson and Goldin-Meadow (2005).

II. Method (Iverson and Goldin-Meadow, 2005)

A complete description of the experiment can be found in Iverson and Goldin-Meadow (2005).

A. Participants

10 participants completed the study, followed longitudinally between the ages of 10 and 24 months. All children were from middle- to upper-middle-class monolingual English-speaking families. Each child was observed an average of 8 times.

B. Procedure

The children were videotaped monthly for approximately 30 minutes. All taping occurred at the child's home during play time with a primary caregiver or during meals. The experimenter brought toys, but children were allowed to play with their own toys as well. Gesture and language were both coded.

Gesture coding

Several types of gestures were coded in this study, but deictic gestures were the focus of this study. Deictic gestures occurred in one of three formats: (1) showing an object by holding up an object to the listener's potential line of sight; (2) index pointing by extending the index finger toward an object; and (3) palm pointing by extending a flat hand toward an

object. In all cases, the referent of a deictic gesture was assumed to be the object pointed at (or held up) by the hand.

Speech coding

All meaningful communicative vocalizations were coded. These vocalizations consisted of either English words or patterns of speech sounds consistently used to refer to a specific object or event (e.g., [ba] for ''bottle'').

Speech+Gesture Coding

All instances in which a child referred to an object were categorized into one of three groups: (1) speech only (using only a word to refer to an object), gesture only (using only a gesture to refer to an object), or speech and gesture (i.e., using both a word and a gesture to refer to an object).

Reliability

Reliability between two independent coders was 93% for gesture and Cohen's kappa was .92 (excellent agreement). Agreement was 100% for assigning meaning to gestures and 91% for assigning meanings to words (no kappa available). Agreement for Speech+Gesture coding was 92%, kappa = .85. (excellent agreement).

C. Results

As Figure 1 suggests, the majority of object referents were made with gestures only, while the number of speech only and speech+gesture were approximately equal. Interestingly, in session 1, 90% of the children had a majority of object references in gesture only, while in the last session, 0% of the children had a majority of references in gesture only.

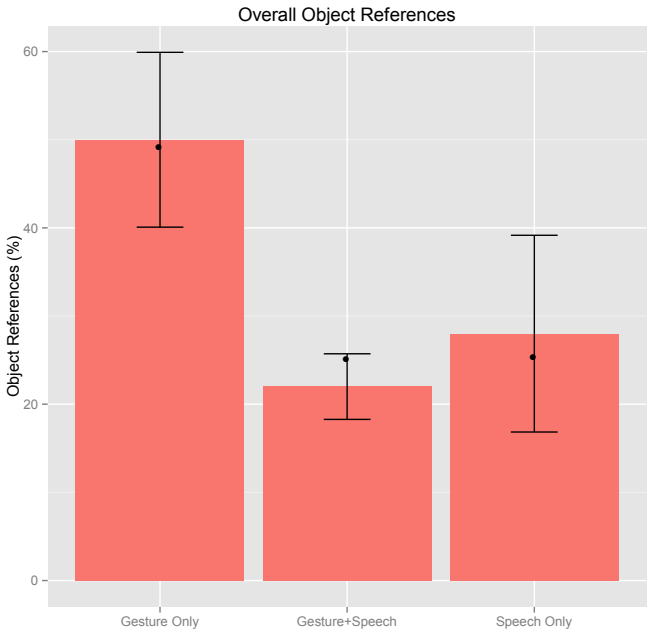

Figure 1. The percentage of object references across all months. The bars are the empirical data [8]**, the points are the model fits, and the error bars are 95% confidence intervals.**

This shift from referring to objects in gesture alone to references using speech or speech+gesture was explored by identifying how specific lexical items were referred to across multiple sessions. There were four possibilities for how words could be referred to across sessions: (1) Speech-Speech (started in speech and stayed in speech), (2) Speech-Gesture (started in speech and transitioned to gesture), (3) Gesture-Gesture (started in gesture and remained in gesture), and (4) Gesture-Speech (started in gesture and transitioned to speech).

As Table 1 suggests, more items were initially referred to in gesture than in speech ($p < .01$). Most importantly, item referents were more likely to move from gesture to speech than from speech to gesture ($p < .001$). On average, children produced a gesture for a specific object 3 months before producing the corresponding word for that object.

Initial Referent	Later Referent	Empirical	Model
Speech	Speech	16% [8-25%]	12%
Speech	Gesture	9% [5-13%]	6%
Gesture	Gesture	25% [17-33%]	38%
Gesture	Speech	50% [42-57%]	44%

Table 1. How lexical items transitioned from one modality to another. Empirical data includes mean and 95% confidence intervals in square brackets.

These results show that a large proportion of object referents appeared in gesture first and then transitioned to language.

D. Discussion

Deictic gesturing seems to precede lexical naming. These results are consistent with the hypothesis that gesture facilitates language development. Iverson and Goldin-Meadow (2005) propose three possibilities for how and why gesture may precede word-level development.

First, gesture may simply be a request from the child that they would like to know the name of a specific object. If a child points to a toy robot, a caregiver may respond with, "Do you want the **robot**?" The caregiver may emphasize the toy name [11] to facilitate the child's symbol grounding. Note that the naming explanation is merely a request for information, not a true facilitation of gesture to language.

Second, gesture may be capitalizing on the fact that gestures frequently use spatial representations and may therefore be used to convey spatial information [2; 3].

Third, gesture may put less demand on memory. Gesture seems to save speakers cognitive effort [12; 13] and it may simply be easier to express a lexical item in gesture than in language.

We next describe the architecture and the task model.

III. Architecture Description

ACT-R/E (Adaptive Character of Thought-Rational/Embodied) is a hybrid symbolic/sub-symbolic production-based system based on ACT-R [14]. ACT-R/E consists of a number of modules, buffers, and a central pattern matcher. Modules contain a relatively specific cognitive faculty associated with a specific region of the brain. For each module, there are one or more buffers that communicate

directly with that module as an interface to the rest of ACT-R/E. At any point in time, there may be at most one item in any individual buffer; thus, the module's job is to decide what and when to put a symbolic object into a buffer. The pattern matcher uses the contents of the buffer to match specific productions.

ACT-R/E interfaces with the outside world through the visual module, the aural module, the motor module, and the vocal module. Other current modules include the intentional, imaginal, temporal and declarative modules. ACT-R/E perceives the physical world by robotic sensors and effectors to it and uses a spatial module [15]. ACT-R/E's goals are to maintain cognitive plausibility as much as possible while providing a functional architecture to explore embodied cognition, cognitive robotics, and human-robot interaction.

Below we highlight the architectural components that are relevant to this project. Figure 2 shows a schematic of ACT-R/E and is discussed more fully in Trafton et al. (2013). [16]

A. Visual

The Visual Module is used to provide a model with information about what can be seen in the current environment. ACT-R/E is able to accept input from video sensors. The visual module allows access to both the location of an object (the "where"' system) and a more detailed representation (the "what" system). Obtaining additional information about an object or person requires declarative retrieval(s). Objects are initially detected using a technique known as Chamfer matching [17]. Chamfer matching uses a shape dictionary to localize and characterize objects in the environment. We do not make a strong plausibility argument for Chamfer matching, though we do notice that both biology and Chamfer rely on dictionary-like representations. Objects are further characterized by their color, derived from the hue of the object.

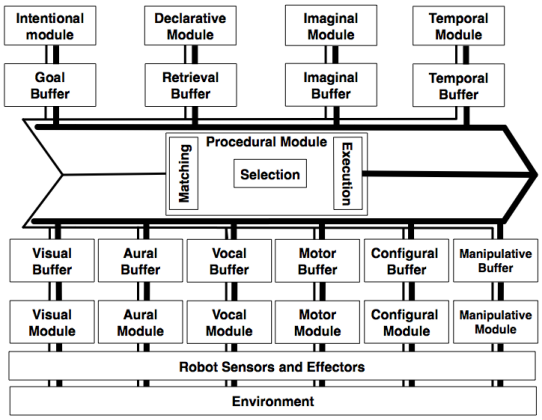

Figure 2: Schematic of ACT-R/E

B. Motor

ACT-R/E's motion is controlled by the motor module. In this project, motor is used to control the robot's eyes and hand for pointing.

C. Spatial

Spatial representation is a critical aspect of being able to point to an object and to be able to pick up an object. ACT-R/E performs pointing by utilizing its spatial representations, collectively known as Specialized Egocentrically Coordinated Spaces (SECS, pronounced seeks) [15]. SECS is neurologically inspired and based on 3D space [18]. SECS provides two egocentric spatial modules, which are responsible for the encoding and transformation of representations in service of navigation (configural) and manipulation and pointing (manipulative).

The configural module provides high fidelity location information for attended representations that is automatically updated as the model moves through or looks around the environment. The configural module represents the world as spatial blobs that can be navigated to or around. The manipulative module uses a metric, geon-base [19] 3D representation for objects. The manipulative module provides encodings of object geometry and orientation and position in support of pointing and grasping, a critical component to object shape description discussed below.

IV. Simulator and Robot Description

Currently, the open-source Gazebo robot simulator [20] is used to enable data collection and to speed-up the model development cycle.

Our current robot platform is the MDS (Mobile-Dexterous-Social) Robot [21]. The MDS robot neck has 18 DoF for the neck and head including eye pitch and pan which allows the robot to look at various locations in 3D space. Perceptual inputs include a color video camera and a SR3000 camera to provide depth information (unused in this project). For the current project, the MDS head can move its eyes and head and its arm and hand and fingers to look at and point at various objects in 3D space.

V. Model Description

An ACT-R/E model was developed that simulates the development of language learning through gesture.

A. High Level Description of the gesture-language model

There are four model components that allow the model to learn single words and receive facilitation from gesture: requesting words through gesture; creating spatial representations of pointed or handled objects; priming of words through visual and spatial representations; and gradual learning of memory representations.

Requesting words through gesture. The model's primary goal is to explore its world. This exploration occurs primarily through playing with objects in the environment. Of course, the model can not always reach or play with an object (due to mobility or manipulation constraints), so it will point at the object. By pointing at an object, the caregiver has a 50% chance of labeling the object. We base that 50% value on Messer's (1981) work, who found that labels in caregivers speech had a 50% chance of being the loudest word in the

utterance. We used this value because the empirical paper idd not report how often the caregiver labeled objects for the child. In simulation, we found that all but the very smallest values of this parameter worked (very low values simply took more time to learn toy names). According to the model, initial gesture to an object or toy served as a request for information.

Spatial representations of pointed or handled objects. According to SECS, any time that an object is grasped or manipulated, a spatial representation is created for that object to enable the manipulation itself. Specifically, the information contains 3D geon [19] shape information (e.g., cube, pyramid, sphere, etc.). Young children do seem to learn both metric and symbolic information about shape at a very early age, so both types of information were included in the spatial representation.[1]

Gradual learning of memory representations. ACT-R/E has a strong theory of memory: how a memory element (chunk) is encoded, remembered, and the situations where forgetting occurs.

Memory in ACT-R/E is described by a chunk's activation. Activation is the log odds that a particular chunk will be useful in the future; high activation chunks are expected to be very useful while low activation chunks are expected to be less useful.

Activation depends both on how much and how frequently a memory has been used in the past, as well as how related the item is to other memories that are currently the focus of attention. Activation consists of three primary components: activation strengthening, spreading activation, and noise. Activation strengthening is learned over time and is a function of how frequently and recently the memory has been thought about in the past, and represents the model's familiarity with a concept. Spreading activation is context dependent, allowing memories that are currently the focus of attention to activate, or prime, other related items. Noise is a random component added in to model the noise of the human brain. They are combined according to the following equation (Anderson, 2007):

$$A_i = B_i + \sum_j W_j S_{ji} + \varepsilon$$

where A_i is the total activation of chunk i, B_i is the total activation of chunk i, $W_j S_{ji}$ is activation spread from item j to item i, and ε is noise. Activation strengthening of a memory item i is calculated according to (Anderson, 2007):

$$AS_i = ln(\sum_{r=1}^{R} t_r^{-d})$$

where R is the number of times item i has been referenced (e.g., was the focus of attention, or was explicitly thought about) in the past, t_r is the time that has passed since the rth reference, and d is the strengthening learning parameter, which defaults to 0.5.

Activation values of 0 or less means that a chunk is not able to be retrieved. However, when two identical chunks are merged, the unitary activation grows with increasing R.

The activation strengthening equation suggests that a new memory element is easy to remember in the short term, but if it is not rehearsed in some manner it will soon be forgotten. The model posits that young children do not have a rehearsal strategy and therefore things will be forgotten quickly. Thus, the model predicts that it will require many repetitions of hearing an object's name before that child is able to remember the name of that object. Of course as a child grows, they are able to create or learn rehearsal strategies so that they are able to remember names of objects and people. The model is able, with nothing but the repetition of a word, to learn the name of objects in its environment over the course of several months.

Priming of words through visual and spatial representations. According to the model, the core reason that gestures facilitate language learning is that aspects of the object representation prime the name of the object. This priming occurs through ACT-R/E's spreading activation, which is part of the declarative memory system.

Spreading activation is spread along associations between memories. In addition to considering what items are being referenced at any given time, it also considers what items are in the current context. The current context consists of both those items being referenced, as well as the set of items in slot values of the items being referenced that are under consideration. Association strengths, intuitively, reflect how strongly item j, when currently being referenced, predicts that item i will be referenced next. The equations for the associative strength from an item j to an item i in memory are

$$S_{ji} = S - \ln(fan_j)$$

where S_{ji} is the strength of association between chunks j and i, S is the maximum associative strength, and fan_j is the fan of chunk j (the number of other memory elements that memory j is associated with).

Every time the model sees an object and wants to pick it up or point to it, a shape representation is created. This shape representation then increases the activation of the name of the object through priming, making the name easier to remember.

For all models, we kept most of the ACT-R/E parameter defaults. The parameters that were changed include the strength of association (ACT-R/E has no default, but we used a common value of 3.5) and activation noise ((ACT-R/E has no default, but we used a common value of .25). All other parameters were set to ACT-R/E default values.

B. Model Environment

The environment that the model lives in contains 30 toys that it is interested in. The number 30 is derived from several empirical papers showing that middle-class families have at least 30 toys for their 1 year old [23; 24]. The toys varied with

[1] Note that this is almost assuredly a simplification. Children do seem to learn spatial symbolic information by 1.5 years [22]. In our work here, we do not model the full learning trajectory, making the assumption that these children have at least some rudimentary spatial knowledge.

respect to shape and color, though there were many overlapping features (e.g., several robots, balls, and dolls, etc.).

C. How the model plays with toys

The model begins at age seven months with very little *a priori* knowledge. It knows the name of no objects and has only a few goals: to find, name, and play with objects.

The model begins by searching the environment for a toy it wishes to play with. The current system has no preferences, so it picks a toy at random from all that are available to it. Once the model has found a toy, it points to it with either a finger or an open hand.

This initial pointing or grasping impacts both the model and the caregiver. If the model wants to know the name of the toy as it is pointing to or holding the toy, the shape and color of the toy provide some activation facilitation to help the child remember the name. The caregiver, seeing the child model point to the toy, may label the toy (e.g., "Do you want the toy **robot**?"). Note that the caregiver provides some prosodic cues to the model [11; 25] to help the child determine the toy name in the speech stream.

The model then plays with the toy for a while. After it is finished playing with the toy, it drops the toy and looks for a new toy to play with. Note that the model may choose the just dropped toy again (though this does not greatly impact the model's performance).

If the model hears the name of the toy the model fuses the aural aspect of the toy (e.g., "ball") with the shape (e.g, sphere) and the color (e.g., red) into a single chunk. This new chunk may be merged with another identical chunk, increasing its activation for later retrieval.

If the model does not hear the name of the toy, an unnamed chunk with the physical characteristics of the toy (e.g., sphere and red) is created. This chunk will exist in memory, but can not provide any information about the name of the chunk. If it is retrieved, the model will not be able to retrieve the name of the chunk, perhaps explaining why children sometimes have a successful retrieval of the object, but can not remember the name itself.

If the model is able to remember the name of the chunk, it will either say the name of the toy or it will say the name of the toy while gesturing to it.

During the first three months of the model's life (age 7 – 10 months), the model spends 30 minutes a day showing a toy to its caregiver and playing with toys. We do not have any strong data on 30 minutes / day, but this number was chosen for several reasons. First, 30 minutes is the length of the experiment itself, and a different number would have increased the number of free parameters to the model. Second, while we are quite sure that most middle-class children play with toys longer than 30 minutes, it is not clear how long each day a caregivers will label toys for the child.

At age 10 m (the average age of the children in the study), a 30 minute experimental session was run. The experimental session is exactly the same as the daily session for the model, but data is collected about what items are labeled by the caregiver and played with by the model.

Data is then collected for 8 months (the average time the experiment lasted during the study) with data collected once a month, but the child playing with toys every day.

D. A sample experimental model run

The first time the model is run in experiment mode, the model has played with many toys and has received labels for many of them from previous interactions with its caregiver. When the experiment begins, the model finds a random toy in its environment that it wants to play with – in this case, a blue box and attempts to retrieve its name. The blue box's representation has this form:

box4293
  isa object
  name "box"
  shape cube
  color blue

Because the box has only been labeled for it a handful of times over the previous three months, the model is unable to retrieve the name of the toy because its activation is -1.97 (recall that chunks can not be retrieved unless they have activation over 0).

The model then points at the box and that pointing creates a spatial representation of the ball (cube) which can then be used to spread activation to the box object chunk. Unfortunately in this case, the total amount of spreading activation is only .75 and noise is a -.28, which is not enough to bring the total activation above 0 and allow the child to remember the box chunk and hence remember its name.

The pointing, however, is noticed by the caregiver and the caregiver says "box." The model then creates a new chunk that contains exactly the same slots and values; this chunk is then merged with the box4293 chunk above which increases the number of references to the named box chunk, increasing the activation for the next time that it is needed. As the object is heard multiple times, the activation gradually increases and eventually the named chunk is able to be retrieved. When the model is able to remember the name, the model will either say the toy's name or say its name while pointing to or handling the toy.

E. Modeling developmental progress

When the model is young, it has no knowledge about object names, what they look like, or any other features about them. As it gets the opportunity to play with different toys and hear their caregiver name the toys, it builds up a simple representation of the toy that includes the name, the shape, and the color of the toy (among other things).

This representation is very weak at the beginning – it is quite difficult to remember since the child's experience with it is so little. However, as the model has more experience with the toy, the memorial representation strengthens and is finally able to be recalled consistently.

Gesture helps in two different ways. First, gesture provides

a cue to the caregiver that the model wants a label for the object. This is especially critical during the early parts of the model's life. Second, gesture provides a source of priming to the object. This priming literally speeds up learning because it increases the activation of the gestured object.

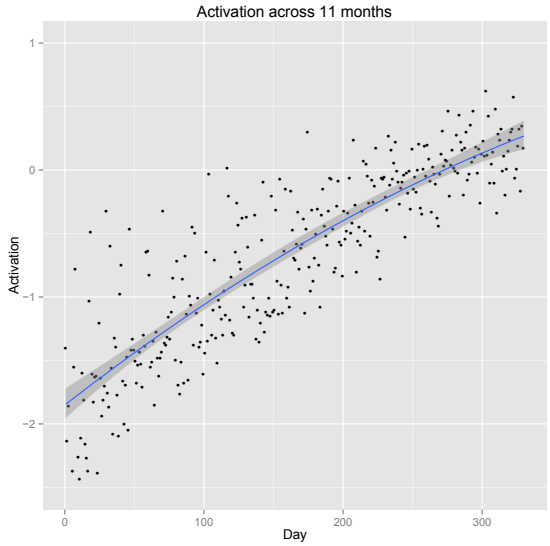

Figure 3: A toy's activation across the entire experiment. For this model, priming was turned on. The black dots are the activation at the end of each day of the model experiment and the line shows a best fitting line with 95% CI.

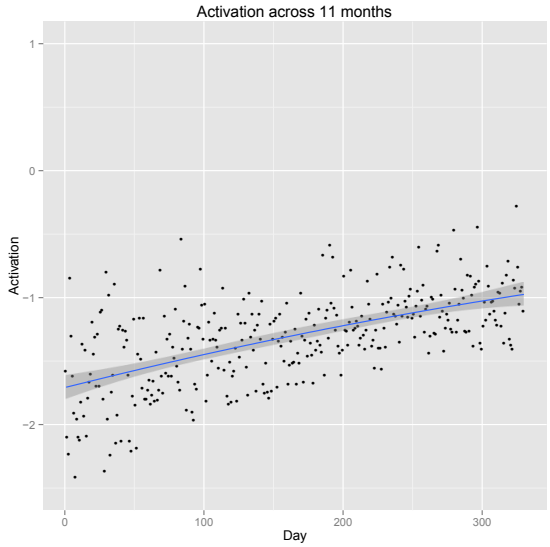

Figure 4: A toy's activation across the entire experiment. For this model, priming was turned off. The black dots are the activation at the end of each day of the model experiment and the line shows a best fitting line with 95% CI.

To explore the advantage of gestures and priming, Figure 3 shows a graph of a toy's activation across days for a single model. For this model, activation priming was turned on. Figure 4, in contrast, shows a similar graph where priming was turned off. Note that learning occurred much faster with priming: the slopes are quite different. Remember that a word is remembered when its activation is over 0.

To provide some match to the experimental procedure, 16 models were run with priming turned on.

F. Model fit

As is evident in Figure 1, the model matches the overall number of object references quite well; $R^2 = .96$ and RMSD = 2.4. Critically, all model points are within 95% confidence intervals of the data. Additionally, the empirical data showed that 10% of the data on the first experimental session was speech; this model shows a comparable finding with 6% of the model data using speech (within 95% CIs). The empirical data showed that 0% of the children used a majority speech on the last session whereas 6% of the model children used a majority of speech on the last session.

Finally, as Table 1 suggests, the transition model data shows a reasonably close fit to the data, with an $R^2 = .79$ and RMSD = 7.5. This data is quite interesting; the current model does transition from gesture to speech, but seems to continue to gesture at a rate slightly slower than the empirical data suggests (the gesture-gesture transition is the only data point outside of confidence intervals).

This model was run using Octavia.

VI. General Discussion

This paper has described a process model of how children use gesture to facilitate language learning. In this model, gesture has two primary roles: a request for the name of the object; and to provide priming to the memorial representation to facilitate memory. When the model begins, it does not know the name of any objects; the only way to get that information is from an external source – in this case a caregiver. When the model originally hears a label for an object, its memory for the object is quite weak and is likely to be forgotten soon. When the model gestures at an object, it also creates a spatial representation that provides some priming activation to the name of the object. As the model hears the name of the object multiple times during its life, its activation strengthens; additional priming helps the model learn the object name faster.

This model connects well with the original empirical paper. Recall that Iverson and Goldin-Meadow (2005) suggested three reasons for how and why gesture precedes word-level development.

They proposed that a child's gesture may tell a caregiver that they want to know the name of an object. In our model, this is a near-necessity, since there is no other way that the model can request the name of an object besides pointing or handling the object. We assume that a child can also extract words from the speech stream, but that gesturing greatly speeds this process up.

Second, Iverson and Goldin-Meadow suggest that gestures may actually convey spatial information. Our model suggests that gestures are a core component of linking spatial representations (shape information in particular), which is a necessary aspect of priming in our model. Thus, while we agree that some gestures can convey spatial information compared to language, our model suggests that this is not where the advantage lies: rather, gesture creates a spatial

representation that provides a boost in activation to a memory chunk and that makes it more likely to remember an object's name.

Third, Iverson and Goldin-Meadow suggest that gesturing reduces cognitive effort by putting less demand on working memory. Our model suggests that priming does not put less demand on memory or save cognitive effort. Rather, it provides a direct facilitative role through spreading activation.

Of the model's four components (requesting words through gesture, spatial representations, priming, and gradual learning of memory representations), three of them are absolutely critical to the success of the model. Getting the names of objects could be done in different ways (e.g., through the speech stream alone). We believe that gesture speeds up this process, but it is not a critical aspect of the success of the model itself. The other three components, however, are needed. Creating spatial representations from gesture is needed because they are a source of priming. Priming itself is the core reason for the facilitation of words. And the gradual learning of the memorial representations is needed to provide a baseline of activation that priming can act upon.

The model also makes a series of interesting predictions. First, it predicts that sometimes, even after a child has successfully named an object, the child may not be able to remember the specific name of an object, but that this should occur less and less as familiarity with a specific object becomes greater. Second, the model shows an interesting pattern where it over-generalizes similarly shaped objects. It can become 'stuck' on some objects, ignoring color or other attributes. Third, the model suggests that it is quite difficult to remember a completely novel object if it is only seen once every month. In fact, the model makes a strong prediction that a novel object can not be learned until a child has the ability to elaborate and / or rehearse the new object. This prediction arises from the activation dynamics described before: an object that is seen and named only rarely does not receive enough base level activation to be able to retrieve its name (this is similar to not remembering an acquaintance's name if you haven't seen them in a very long time).

Overall, this model provides a process explanation for why children gesture to objects and how a gesture can facilitate word-learning.

ACKNOWLEDGMENT

This work was supported by the Office of Naval Research and OSD to JGT.

REFERENCES

1. Goldin-Meadow, S. 2005 Hearing gesture: How our hands help us think. Harvard University Press.
2. McNeill, D. 1992 Hand and mind: What gestures reveal about thought. University of Chicago Press.
3. Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L. D., Schunn, C. D., and Kirschenbaum, S. S. 2006. The relationship between spatial transformations and iconic gestures. Spatial cognition and computation. 6, 1, 1-29.
4. Özçalışkan, Ş. and Goldin-Meadow, S. 2005. Gesture is at the cutting edge of early language development. Cognition. 96, 3, B101-B113.
5. Sauppé, A. and Mutlu, B. 2014. Robot deictics: How gesture and context shape referential communication. Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, 342-349.
6. Breazeal, C. and Scassellati, B. 2002. Robots that imitate humans. Trends in cognitive sciences. 6, 11, 481-487.
7. Scassellati, B. 1999 Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In Computation for metaphors, analogy, and agents, Springer.
8. Iverson, J. M. and Goldin-Meadow, S. 2005. Gesture paves the way for language development. Psychological science. 16, 5, 367-371.
9. Sheldon, M. and Lee, M. 2010. A developmental approach to the emergence of communication in socially situated embodied agents. Development and Learning (ICDL), 2010 IEEE 9th International Conference on, ICDL, 204-210.
10. Sheldon, M. and Lee, M. 2011. PSchema: A developmental schema learning framework for embodied agents. Development and Learning (ICDL), 2011 IEEE International Conference on 2, ICDL, 1-7.
11. Messer, D. J. 1981. The identification of names in maternal speech to infants. Journal of Psycholinguistic Research. 10, 1, 69-77.
12. Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., and Wagner, S. M. 2001. Explaining math: Gesturing lightens the load. Psychological Science. 12, 6, 516-522.
13. Wagner, S. M., Nusbaum, H., and Goldin-Meadow, S. 2004. Probing the mental representation of gesture: Is handwaving spatial? Journal of Memory and Language. 50, 395-407.
14. Anderson, J. R. 2007 How Can the Human Mind Occur in the Physical Universe? Oxford University Press, USA.
15. Trafton, J. G. and Harrison, A. M. 2011. Embodied spatial cognition. Topics in cognitive science. 3, 4, 686-706.
16. Trafton, J. G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., and Schultz, A. 2013. ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction. Journal of Human-Robot Interaction. 2, 1, 30-55.
17. Liu, M.-Y., Tuzel, O., Veeraraghavan, A., and Chellappa, R. 2010. Fast directional chamfer matching. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, CVPR, 1696-1703.
18. Previc, F. H. 1998. The neuropsychology of 3-D space. Psychological Bulletin. 124, 2, 123-164.
19. Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. Psychological review. 94, 2, 115.
20. Koenig, N. and Howard, A. 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator. Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on 3, IROS 2004, 2149-2154.
21. Breazeal, C. 2009. MDS Robot.
22. Smith, L. B. 2009. From fragments to geometric shape changes in visual object recognition between 18 and 24

months. Current Directions in Psychological Science. 18, 5, 290-294.
23. Power, T. G. and Chapieski, M. L. 1986. Childrearing and impulse control in toddlers: A naturalistic investigation. Developmental Psychology. 22, 2, 271.
24. Rheingold, H. L. and Cook, K. V. 1975. The contents of boys' and girls' rooms as an index of parents' behavior. Child development. 459-463.
25. Kriz, S., Anderson, G., Bugajska, M., and Trafton, J. G. 2009 Talking to Robots: Features of Robot-Directed Speech. In Human Robot Interaction 2009,