Acquisition of Lexical Semantics through Unsupervised Discovery of Associations between Perceptual Symbols

Tuna Oezer
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA

***Abstract*—This paper introduces an unsupervised method to acquire the lexical semantics of action verbs. The eventual goal of the presented method is allowing a robot to acquire language under realistic conditions. The method acquires lexical semantics by forming association sets that contain general perceptual symbols associated with a certain concept as well as perceptual symbols of the utterances of the name of a concept. The lexical semantics is learned with the help of a narrator who comments on what the robot sees. The technique works even if the narrator only occasionally comments on what the robot sees. The paper presents experimental results that show that the method can acquire the lexical semantics of action verbs while the robot is watching a human who performs actions and hearing a narration that only occasionally actually describes what the robot is currently seeing. A comparison with supervised learning algorithms shows that the method discussed in this paper outperforms other techniques.**

Keywords-language acquisition; lexical semantics; perceptual symbol systems; unsupervised learning

I. INTRODUCTION

Children have a remarkable ability to acquire language. It has been a challenge to replicate this ability in machines. The research discussed in this paper uses a developmental approach to tackle this problem. This paper presents an unsupervised algorithm for acquiring lexical semantics that was designed to work in a robot that mimics a child's development. In particular, the input to the robot is very similar to the type of input a small child would receive, i.e. sensory input combined with the speech of adults.

This research focuses solely on semantics and ignores syntax. In particular, the experiments presented in this paper deal with the acquisition of lexical semantics of certain action verbs. However, the presented method is also applicable to general lexical semantics acquisition tasks.

Children can acquire language rather quickly in part because they usually have some understanding of the meaning of the words they learn. In particular, when children start to acquire language, they already have a basic world model. Thus, language learning must be preceded by a stage in which the child acquires a basic model of its environment.

One important problem is the particular representation of the basic world model. The research presented in this paper uses a *perceptual symbol system* to represent the world model. This type of representation has been proposed by Barsalou [1]. The perceptual symbols are acquired through sensory-motor interactions with the environment. The method described in this research first acquires a set of perceptual symbols and then acquires lexical semantics using the perceptual symbol system from the first stage as the basic world model. The next sections will provide more details about the nature of the perceptual symbols, how they are acquired and finally how they can be used to learn language.

While the technique discussed in this paper can be applied to more general lexical semantics acquisition problems, the experiments and results presented in this paper focus on the acquisition of the semantics of action verbs. In particular, the results show that the perceptual symbol system acquired by the robot can be used to successfully acquire the semantics of the action verbs in an unsupervised manner even if the robot is presented with narrations that contain many random words.

II. RELATED RESEARCH

This research is most closely associated with autonomous mental development (AMD) and developmental robotics. Ref. [2] gives a brief introduction into AMD. A more detailed overview with some research issues is presented in [3]. An overview of developmental robotics can be found in [4]. A summary of current research in epigenetic robotics, which is related to developmental robotics, is given in [5].

The method discussed in this paper relies on information theoretic measures to identify associations. Information theoretic measures have been successfully used in the past to learn lexical semantics. As an example, Roy and Pentald [6] have developed a computational model called CELL which acquires words from multimodal sensory input. The CELL system uses mutual information to associate utterances with visual input. Unlike the CELL system, this research focuses on actions and deals with sentences which may or may not correspond to what the robot sees. Gold, Doniec and Scassellati present the word trees method, which reconstructs the speaker's decision process in choosing a word [7]. They use

entropy to learn word trees. Yu and Ballard present a system that extracts perceptual representations from sensory data using clustering [8]. Their system uses these representations to ground the meaning of nouns and verbs. Unlike Yu and Ballard's system, the method presented in this paper also allows the narrator to speak sentences unrelated to the video input.

There have been a few projects that have investigated language learning using robots. Ref. [9] presents a variety of methods to train a robot. They teach the robot simple verb-noun commands using supervised learning. Ref. [10] describes a system that learns two word verb-noun sentences. Their system uses separate linguistics and behavior modules that are linked together with a parametric binding method. The MirrorBot robot [11] tries to mimic mirror neurons. A self organizing map is used to map action words to body areas that are used to execute the action, thus replicating neuroscience data. Ref. [12] describes an architecture to acquire language with an autonomous robot that interacts with its environment.

Ref. [13] describes a simulated system that learns single verbs and thus focuses on lexical semantics. In particular, the system maps verb meaning to predefined action schemas called x-schemas. Ref. [14] proposes an implementation of perceptual symbol systems that captures the temporal structure of an action using a recurrent Neural Network. Ref. [15] presents a simulated robot that learns the names of actions. In contrast to the perceptual method presented here, Siskind uses event-logic expressions to recognize the occurrence of spatial motion verbs in short image sequences [16].

There is plenty of neuroscience support for perceptual symbol systems [1]. [17] have conducted a number of brain imaging studies in which subjects where listening to words. The results clearly show that action verbs activate motor neurons which are involved in executing the meaning of the action verb. [18] provide even more evidence that natural language understanding activates motor areas in the brain. [19] were able to show that sentences which contain action words activate a mirror neuron that is involved in processing the action represented by the word. [20] have developed a model for language acquisition in children and have determined that when a predefined sensory-motor model is employed not only is the acquisition of single words simplified, but the acquisition of syntax can be accomplished more easily. Cohen, Morrison and Cannon [21] show that preschool-age children choose words to describe movies based on dynamical aspects of the movies.

The problem discussed in this paper is a more complex variant of the symbol grounding problem [22] (see [23] for a more recent overview). In the method discussed in this paper, the learner has to discover the symbols. The problem is further complicated by the weak association between a word and the visual input it describes, i.e. the word is only sometimes present when the corresponding visual input is present and many unrelated words may be present, too.

III. Association Sets

The method discussed in this paper uses *association sets* to represent meaning. An association set is simply a set of highly associated perceptual symbols.

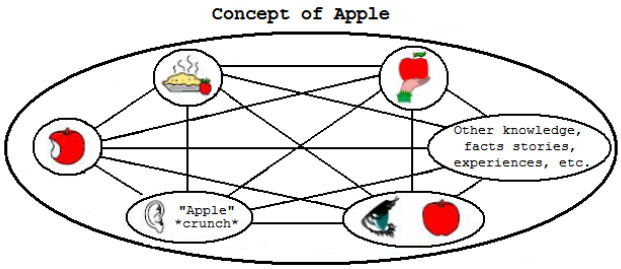

Figure 1. An apple is represented by an association set that consists of perceptual symbols.

An example of such an association set is shown in fig. 1. This figure shows how the meaning of an apple is represented. There is actually no internal symbol named apple. Rather the meaning of an apple is represented through a number of perceptual symbols that describe the visual appearance of apples, the smell of apples, the taste of apples, and so on. These perceptual symbols are linked together in an association set. This particular set of links represents the meaning of an apple.

The set of all association sets functions as the world model. The world model is learned by discovering new association sets. Association sets are formed whenever a strong association between a set of perceptual symbols is identified. In order to be able to identify these associations, the robot needs to interact with its environment.

This type of representation has several advantages. First, the meaning is composed of only perceptual symbols and does not require any built-in symbols that would need to be magically defined by an expert. Second, the meaning of a concept is easily extensible. By adding new links to other perceptual symbols, the meaning of a concept can be expanded or modified. Third, it is possible to compare two concepts by comparing the perceptual symbols in the association sets.

An association set allows the retrieval of all perceptual symbols in the association set given any subset of the perceptual symbols. For example, it is possible to identify all concepts that are associated with a certain sound. This comes in handy in the acquisition of lexical semantics. The utterances of a word are also treated as perceptual symbols. The name of a particular concept can be learned simply by placing the perceptual symbol that represents the utterance of the name of the concept into the association set that represents the concept. This is demonstrated in fig. 1. The utterance “apple” is one of the perceptual symbols in the apple association set.

A key advantage of acquiring lexical semantics via this method is that the *same* algorithm that is used to discover the general association sets can be used to acquire lexical semantics. Initially, association sets will first only associate basic perceptual symbols with each other and will not contain any language utterances. Once the association sets start to mature, utterances will be organically added to the association sets by just continuing the same learning method.

In principal, it is also possible to build a hierarchy of association sets by linking association sets with other association sets. However, this is not further pursued in this paper.

IV. PROBLEM DEFINITION

This section provides a more formal problem definition. Section A discusses the input, section B discusses the required learning task and section C discusses the expected output.

A. *Input*

The input is similar to the type of input a young child would experience. In the experiments discussed in the paper, the input is provided by a conventional video camera and a microphone. The general algorithm is agnostic about the specific type of input such that in general other sensors could be used as well. For the verb acquisition discussed in this paper, a stationary camera was sufficient. However, in general a robot that interacts with its environment can produce richer association sets and thus acquire a better world model. Since the specific focus in the experiments is on action words, the video camera was pointed at a scene in which a human executed a number of actions on a set of objects. In particular, the human carried out one of the following 13 actions: bounce, carry, drop, juggle, kick, lift, lower, pull, push, roll, swing, throw or wave. The actions were carried out with one of the following 6 objects: a ball, a tray, a bottle, a box, a bag, or a chair. In order not to add too much complexity to the vision part, the video was recorded under controlled conditions in a lab.

It is important to point out that there is nothing special about this particular set of actions or objects. The basic algorithm is completely agnostic about its input and can be applied to other setups as well. This setup has been selected primarily to test the algorithm on a reasonably challenging case.

While the human is performing the action, a single narrator utters some comments on the microphone. In principal, the comments can be arbitrary sentences. However, since syntax is not an issue, the comments consist of a set of words. In particular the narrator utters K many words during each sample selected from a list of W many words. With probability Q, one of the K words may be the name of the action that is performed by the actor.

The input to the algorithm is the combination of video and audio. A sample input is shown in fig. 2. For simplicity, the input is provided in the form of samples, where each sample consists of a video sequence depicting the action performed by the human and some audio that contains the narration provided by the narrator.

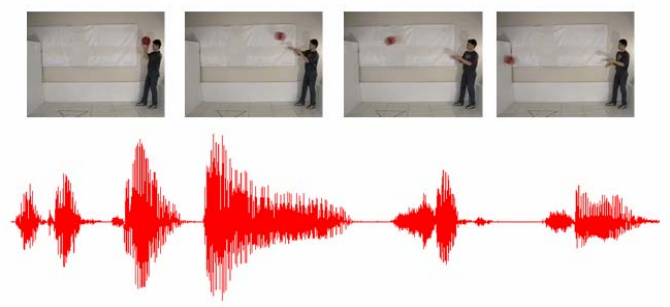

Figure 2. One input sample consists of a video sequence and an audio narration.

B. *Learning Task*

Given the video sequences and narration discussed in section A, the algorithm has to be able to acquire lexical semantics. The number of actions and words in the language are not supplied to the algorithm. First, the algorithm has to identify perceptual symbols in the video and audio. Next, the algorithm has to detect strong associations between the perceptual symbols and identify appropriate association sets. Assuming the algorithm functions as expected, the lexical semantics should be acquired automatically, since the perceptual symbols that represent the utterances associated with the names of the actions will be placed in the same association sets as the perceptual symbols that represent the visual appearance of the action.

As discussed earlier, each narration consists of K words. The K words are selected from a set of W words. With probability Q, one of the K words might be the name of the action. The narrator has no obligation to actually name the action in the narration. For example, if the human kicks a ball, the narrator might comment about pushing a chair. However, it is assumed that the narrator will occasionally comment about the performance of the actor using the actual name of the action.

The weak relationship between the narration and the visual perception provide a much more realistic problem setup. After all, adults do not necessarily always comment about what children pay attention to. In the case of a robot that acquires language from a narrator, the narrator might not always comment on what the robot currently pays attention to, such that a certain tolerance to unrelated comments is required. However, this setup also makes the learning problem much more difficult. Experiments show that many traditional learning techniques fail under these circumstances. The problem is that there is no direct relationship between the utterances and the visual input. In summary, the learning task is quite challenging because:

- The algorithm does not know the number of actions or words in the language.
- The comments provided by the narrator are not guaranteed to include the name of the action visible in the video.
- There is no supervised input that labels the actions or words.

C. *Expected Output*

In essence, the algorithm generates a world model in form of a set of association sets. The algorithm first identifies and generates a list of perceptual symbols. Using this list of perceptual symbols, the algorithm then discovers and outputs a set of association sets.

If the algorithm has functioned correctly, the output should contain one association set for each concept. In this case, there should be precisely 13 association sets, one for each action. Furthermore, each association set should contain visual and auditory perceptual symbols that are associated with only one action.

The algorithm will demonstrate that it has acquired the lexical semantics correctly by placing the perceptual symbol of the utterance of an action in the association set of that action. In other words, given a video sequence of an action, the output of the algorithm could be used to name the action, or vice versa given the name of an action and a long video, a video sequence that depicts the action could be identified.

It is important to understand that the algorithm is unsupervised and thus there is no simple right or wrong answer. The output of this algorithm is more comparable to the output of a self-organizing map. There are many possible combinations of association sets that can be identified all of which are equally correct.

V. ALGORITHM

The algorithm consists of five main steps:

1. Preprocessing
2. Computation of similarity matrices
3. Discovery of perceptual symbols via clustering
4. Computation of the association matrix
5. Discovery of association sets

Steps 3 through 5 are completely agnostic about the specific nature of the input. Steps 1 and 2 have been adapted to the specific problem setup. In order to apply the algorithm to another problem setup, the first two steps may need to be modified. The algorithm is discussed in more detail in [24].

A. Preprocessing

The input consists of a series of samples, each of which consists of some video sequence and some narration. The preprocessing step prepares the input for the similarity computation step. In particular, the visual preprocessing identifies the path along which the object travels and computes the elevation of the object as well as the distance to the human during each step. This requires algorithms to detect the human and object as well as to track the object across frames.

The auditory preprocessing computes several power spectra at different resolutions. The frequency bands of the spectra widen with increasing frequency. The spectra are normalized along each time slice in order to highlight the relative differences between the frequency bands at each point in time.

B. Computation of the Similarity Matrices

The discovery of perceptual symbols requires the ability to determine the similarity between each pair of samples. This is simplified by computing a similarity matrix. The entry in column i and row j of the matrix specifies the similarity between sample i and sample j. One similarity matrix is computed for each modality. The separate similarity matrices allow the discovery of unimodal perceptual symbols for each modality.

The visual similarity matrix is primarily based on the similarity of the object path extracted from each sample. The paths are segmented into a series of straight lines. Two paths are compared by comparing the length, slope, base, average and height of each corresponding pair of line segments. In addition, the algorithm also determines whether an object is pushed or pulled.

In the case of the audio, the similarity matrix is computed by comparing the preprocessed power spectra. It turns out, that after preprocessing an element by element comparison with a time shift is sufficient. Various time shifts are tried to find the best match between the spectra. The distance computation penalizes pairs of power spectra based on the difference of their length.

C. Discovery of Perceptual Symbols

Perceptual symbols are discovered via agglomerative clustering. Each cluster represents one perceptual symbol. Agglomerative clustering starts by placing each sample into a separate cluster and then repeatedly merges nearby clusters with each other. In order to identify nearby clusters, the agglomerative clustering algorithm only needs to know the similarity between each pair of samples. This can be easily looked up in the similarity matrices computed in step 2. The distance between two clusters is determined by the average similarity between all pairs of samples in the two clusters.

The agglomerative clustering stops once all clusters have a minimum size and distance to each other. These two thresholds need to be manually specified. In the experiments discussed in this paper, the agglomeration stops once visual clusters contain at least 30 samples and have a minimum distance to each other of 0.15 based on a maximum distance of 1.0. Audio clusters are required to contain at least 10 samples and have a minimum distance to each other of 0.3.

The set of clusters produced by the third step represents the perceptual symbols. In general, the perceptual symbols do not coincide with any particular high level concepts, but represent repetitive perceptual patterns. For example, a particular utterance may be represented by multiple perceptual symbols.

D. Computation of the Association Matrix

The fourth step determines the strength of association between each pair of perceptual symbols. This information is needed by the last step. The associations are organized in an association matrix. The entry in column i and row j of the association matrix specifies the association between perceptual symbol i and j.

Associations are determined by computing the pointwise mutual information (PMI) between two perceptual symbols. Thus, given two perceptual symbols **a** and **b**, the association between the two symbols is determined via (1).

$$assoc(a,b) = \log \frac{P(a,b)}{P(a)P(b)} \tag{1}$$

In the particular case discussed in this paper, **a** is a visual perceptual symbol and **b** is an auditory perceptual symbol. The association matrix is normalized such that the mean association is zero and the standard deviation is 1.

E. *Discovery of Association Sets*

The association sets can be discovered using the association matrix. Essentially, perceptual symbols that are highly associated with each other are placed into one association set. The discovery of association sets is similar to agglomerative clustering. First, all pairs of perceptual symbols that are highly associated with each other are placed in some association set. Next, pairs of association sets that contain perceptual symbols that are on average highly associated with each other are merged together. This continues until all such association sets have been merged together.

A pair of perceptual symbols is considered to be highly associated with each other if their association with each other is at least 2 standard deviations above the mean. This is an arbitrary number that works well in a number of cases and can be changed if needed.

The final output is a list of association sets. Ideally, each association set represents one concept, which is in this case an action. Thus, all visual and auditory perceptual symbols of a given action should be present in the same association set.

VI. Experimental Results

The algorithm was thoroughly tested with a number of parameter configurations. This section highlights the main results. The experiments were run on a total of 3,074 samples.

A. *Measuring Performance*

Since the algorithm is unsupervised, it is not trivial to measure the performance. There are two criteria that determine a good set of association sets. First, each association set should contain only perceptual symbols that belong to one action. This is a measure of the purity of the association sets. Second, there should be exactly one association set for each action. This is a measure of compactness. The algorithm could achieve perfect purity by producing trivial association sets for each action by generating one association set for each perceptual symbol. Thus, it is desired that the output is as compact as possible and as pure as possible.

This is conveniently measured by the entropy product. The entropy product is the product of two entropy measures. The first entropy measure reflects the purity of the association sets and the second measure reflects the compactness. For simplicity, entropy product values are normalized such that the best value is 1 and the worst is 0.

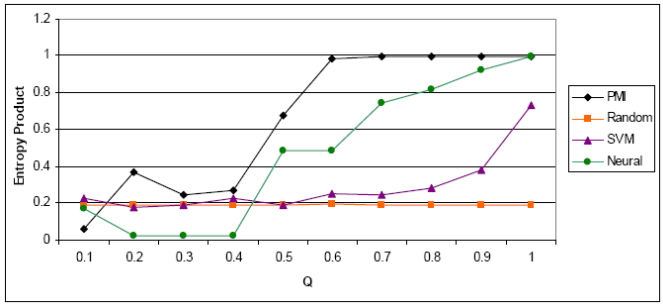

Figure 3. Experimental results. PMI is the method discussed in this paper.

The full formula for the entropy product is quite lengthy and is presented in [24]. In summary, if all association sets contain only perceptual symbols of one action and there is exactly one association set per action, the entropy product will be 1. Otherwise, if a few association sets contain perceptual symbols of more than one action or there is more than one association set for some actions, the entropy product will be less than 1. If all association sets contain all perceptual symbols, the entropy product will be zero.

The entropy product is determined under a number of parameter configurations. The most important one of these parameters, called Q, specifies the probability that one of the K words in the narration is the actual name of the action. Thus, if Q is 1, the narrator mentions the name of the action among the K words every time the actor performs the action in the video. This guarantee makes the problem relatively simple. If Q is below 1, given N samples approximately only Q × N many samples will contain a narration that mentions the name of the action. In the remaining samples, the narration will be completely unrelated to the action. Thus, as the Q value is lowered, the problem becomes more challenging. Traditional algorithms will typically fail if the Q value is significantly lowered.

B. *Main Results*

The main result is summarized in fig. 3. This figure shows the results when W, the number of words in the language, is 30 and K, the number of words in each narration, is 5. As can be seen, the technique proposed in this paper and indicated with the label PMI (pointwise mutual information) in the figure performs very well until Q drops to a value of about 0.5.

In order to be able to better evaluate the results, the algorithm has been also compared with two supervised learning algorithms. These two algorithms are support vector machines (SVM) and neural networks. Note that unlike the method discussed in this paper, the supervised learning algorithms receive the true labels of the samples. The SVM has been trained with a linear kernel and the neural network is a standard 3 layer network that was trained with back propagation.

As can be seen from fig. 3, the technique discussed in this paper outperforms the two supervised learning algorithms. In particular, the supervised learning algorithms have trouble for Q values below 1, which is not surprising. Supervised learning algorithms are not designed for that type of learning problem. In contrast, the method presented in this paper has been designed precisely for such conditions.

Other experiments show that the performance of the algorithm decreases as K decreases, while the performance increases as W increases. In fact it can be shown that the strength of association between perceptual symbols linked to a given concept is proportional to (2).

$$\log \frac{Q \times W}{K} \tag{2}$$

Thus, it is easier to discover association sets if the language contains many words. In contrast, it gets more difficult if the narrator speaks in long sentences.

Other experiments show that the algorithm can very reliably distinguish function words from content words. Furthermore, the algorithm is very robust under noisy conditions.

The algorithm presented in this paper has been also tested under a variety of other parameter configurations and with other data sets. In all cases, it has shown robust results and has outperformed the two supervised learning algorithms. A more complete list of results is provided in [24].

VII. Conclusion

This paper has presented a technique that is capable of acquiring lexical semantics in an unsupervised manner by discovering associations between perceptual symbols. The technique even works if the narration heard by the robot is only occasionally related to what the robot sees. The method presented in this paper outperforms standard supervised learning algorithms.

The technique discussed in this paper has several advantages. First, it is relatively simple. As a consequence, the algorithm is quite fast and scales well to large sample sizes. Furthermore, its simplicity makes it more biologically plausible. Second, the use of perceptual symbols and association sets make the algorithm more easily extensible to other problems. The method can be easily applied to other problems, since the core parts the algorithm are agnostic about the specific input.

The method presented in this paper is a very promising direction to achieve general language acquisition in robots. While this paper has focused on action verbs, future versions of the algorithm may also support more complex words and sentences. The level of complexity that can be supported primarily depends on the underlying perceptual symbols and how well they can be acquired. A robot that can interact with its environment in a rich way would be able to acquire more complex perceptual symbols. Furthermore, future versions of the algorithm may also provide a mechanism to hierarchically build association sets from more primitive association sets opening up the possibility to represent higher level semantics.

References

[1] L. W. Barsalou. Perceptual Symbol Systems. Behavioral and Brain Sciences, Vol. 22, pp. 577-660. Cambridge University Press. Cambrdige, MA, 1999.

[2] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. "Autonomous Mental Development by Robots and Animals". Science Magazine, Vol. 291, No. 5504, Issue 26, pp. 599-600. American Association for the Advancement of Science, 2001.

[3] J. Weng. A Theory for Mentally Developing Robots. Proceedings of the 2nd International Conference on Development and Learning. IEEE, 2002.

[4] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental Robotics: A Survey. Connection Science, Vol. 15, No. 4, pp. 151-190. Taylor & Francis Group, 2003.

[5] L. Berthouze and C. G. Prince. Introduction: The Third International Conference on Epigenetic Robotics. Proceedings of the 3rd International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Lund University Cognitive Studies, Vol. 101, Lund, Sweden, 2003.

[6] D. K. Roy and A. P. Pentland. Learning Words from Sights and Sounds: a Computational Model. Cognitive Science, Vol. 26, pp. 113 - 146. 2002.

[7] K. Gold, M. Doniec, and B. Scassellati. Learning Grounded Semantics With Word Trees: Prepositions and Pronouns. Proceedings of the 6th International Conference on Development and Learning, London, UK. 2007.

[8] C. Yu and D. H. Ballard. A Multimodal Interface for Grounding Spoken Language in Sensory Perceptions. ACM Transactions on Applied Perceptions, Vol. 1, No. 1, pp. 57 - 80. 2004.

[9] Y. Zhang and J. Weng. Grounded Auditory Development by a Developmental robot. Proceedings of International Joint Conference on Neural Networks, Vol. 2, pp. 1059-1064. Washington, DC, 2001.

[10] Y. Sugita and J. Tani. A Connectionist Approach to Learn Association between Sentences and Behavioral Patterns of a Robot. Proccedings of 8th International Conference on Simulation of Adaptive Behavior, pp. 467-476. Los Angeles, CA, 2004.

[11] S. Wermter and M. Elshaw. Learning Robot Actions Based on Self-organising Language Memory. Neural Networks, Vol. 16, pp. 691-699. Pergamon, 2003.

[12] S. E. Levinson, K. Squire, R. S. Lin, and M. McClain. Automatic Language Acquisition by an Autonomous Robot. AAAI Spring Symposium on Developmental Robotics, 2005.

[13] D. R., Bailey, J. A. Feldman, S. Narayanan, and G. Lakoff. Modeling Embodied Lexical Development. In Proceedings of the 19th Annual Meeting of the Cognitive Science Society, pp. 19-24. Stanford University Press, Stanford, CA, 1997.

[14] D. W. Joyce, L. V. Richards, A. Cangelosi, and K. R. Coverntry. On the Foundations of Perceptual Symbol Systems: Specifying Embodied Representations via Connectionism. In Detje, F., Dorner, D., and Schaub, H. (eds.), The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modeling, pp. 147-157. Universitatsverlag Bamberg, Germany, 2003.

[15] A. Cangelosi and T. Riga. An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments with Epigenetic Robots. Cognitive Science, 2006, 30(4), pp. 673-689.

[16] J Siskind. Grounding Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. Journal of AI Research, Vol. 15, pp. 31 - 99. 2001.

[17] F. Pulvermüller, M. Härle, and F. Hummel. Walking or Talking?: Behavioral and Neurophysiological Correlates of Action Verb Processing. Brain and Language, Vol. 78, pp. 143-168. Academic Press, 2001.

[18] B. K. Bergen and K. B. Wheeler. Sentence Understanding Engages Motor Processes. Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society, 2005.

[19] G. Buccino, L. Riggio, G. Melli, F. Binkofski, V. Gallese, and G. Rizzolatti. Listening to Action-related Sentences Modulates the Activity of the Motor System: A Combined TMS and Behavioral Study. Cognitive Brain Research, Vol. 24, pp. 355-363. Elsevier, 2005.

[20] S. R. Howell, D. Jankowicz, and S. Becker. A Model of Grounded Language Acquisition: Sensorimotor Features Improve Lexical and Grammatical Learning. Journal of Memory and Language, Vol. 53, pp. 258-276. Elsevier, 2005.

[21] P. R. Cohen, C. T. Morrison, and E. Cannon. Maps for Verbs: The Relation between Interaction Dynamics and Verb Use. In Proceedings of the 19th International Conference on Artificial Intelligence. 2005.

[22] S. Harnad. The Symbol Grounding Problem. Physica D, 1990, Vol. 42, pp. 335-346, 1990.

[23] M. Taddeo and L. Floridi. Solving the Symbol Grounding Problem: a Critical Review of Fifteen Years of Research. Journal of Experimental and Theoretical Artificial Intelligence, 2005.

[24] T. Oezer, "Discovering Audio-Visual Associations in Narrated Videos of Human Activities," Ph.D. dissertation, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign, Urbana, IL, 2008.