

Cross-situational noun and adjective learning in an interactive scenario

Yuxin Chen, David Filliat

▶ **To cite this version:**

Yuxin Chen, David Filliat. Cross-situational noun and adjective learning in an interactive scenario. Joint IEEE International Conference Developmental Learning and Epigenetic Robotics (ICDL-EPIROB), Aug 2015, Providence, United States. hal-01170674

HAL Id: hal-01170674
https://hal.science/hal-01170674

Submitted on 18 Aug 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cross-situational noun and adjective learning in an interactive scenario

Yuxin Chen David Filliat
ENSTA ParisTech - INRIA FLOWERS team
Computer Science and System Engineering Laboratory
ENSTA ParisTech
Palaiseau, France

***Abstract*—Learning word meanings during natural interaction with a human faces noise and ambiguity that can be solved by analysing regularities across different situations. We propose a model of this cross-situational learning capacity and apply it to learning nouns and adjectives from noisy and ambiguous speeches and continuous visual input. This model uses two different strategy: a statistical filtering to remove noise in the speech part and the Non Negative Matrix Factorization algorithm to discover word-meaning in the visual domain. We present experiments on learning object names and color names showing the performance of the model in real interactions with humans, dealing in particular with strong noise in the speech recognition.**

I. Introduction

Future robots that will act at home will need a high degree of adaptability to cope with the very complex environment represented by the common household settings. Among many capabilities, they will require an efficient object recognition capacity and the ability to interact with humans about these objects through natural dialogue. This will require learning not only the names of objects, but also some of their features such as size or color. These capabilities can be implemented in order to cope with a variety of situations, but it is impossible to implement in advance the recognition capabilities for all possible objects. Therefore these robots will need to be able to learn new objects, and possibly new words.

The problem faced is the problem of symbol grounding [1], i.e. learning the association of a word to its meaning. It is possible to resort to specific training scenarios for teaching objects names, for example by placing the objects in specific situations and associating them with a label. In such settings, the problem would be cast as a supervised learning problem. However, it would be better to be able to learn in less constrained scenarios during natural human-robot interactions. In such a setup, the problem is more complex as there could be ambiguities both on the object of interest, when several objects are present in the field of view of the robot, and on the labels to be associated with it as the human could pronounce complex sentences where several words could correspond to the label to learn. This setup will cast the problem as a weakly supervised or un-supervised problem.

An additional difficulty in the case of language, compared to simple label associations, is that words have different functions and relations to the objects. Lets take a simple example : "Look at this red ball". This sentence contains words and a verb which are not related to the object identity. It also contains a noun and an adjective that are related to the object type (a ball) and to a feature of this object (its color, red). In this situation, it should therefore be learned that "ball" is labelling the object identity, while "red" refers to its color and that all the other words are irrelevant.

Humans are naturally facing these situations every day. It has been demonstrated that as young as 12 months old, we rely on cross-situational learning [2] in order to solve these ambiguities. The general idea is that by analysing the common points between several situations displaying various objects and various associated words, it is possible to solve the ambiguities and to recover the correct object-noun or feature-adjective associations. Several models of cross-situational learning have been proposed, working on symbolic or sub-symbolic representation, with or without noise.

In this paper, we propose an implementation of cross-situational learning which is a first step towards a system for learning objects and words on a robot during natural interaction with humans. The model we propose is able to use symbolic and continuous data that are both ambiguous and noisy in order to learn simple word-meanings in the visual domain. It is implemented and evaluated during real interactions with humans.

II. Related work

Learning the association between words and their meanings is fundamentally ambiguous as illustrated by Quine with the "Gavagai" problem [3]. In this problem, the word "Gavagai" is pronounced while pointing to a rabbit in a field, and therefore its meaning can be "rabbit", "field", or even the color of the rabbit. These ambiguities in communication can be reduced by several method, for example by relying on joint attention [4] or on the syntactic constraints of the language itself [5]. However, human infants and adults are able to learn word meanings even in controlled ambiguous situations by using the regularities across exposure to the words and their referents [6], [7], [8].

An early rule-based model of this capacity has been proposed by [6]. This model is working on symbolic representation and is only evaluated in simulation. More recently, two general approaches have been proposed to model the human

performance [9]. One approach is based on hypothesis testing and assumes that human retains only one association hypothesis when a word appear, which is subsequently confirmed or not [10]. The other approach is based on associative learning and memorizes an association table between all words and referents that is updated after each observation [9]. Empirical evaluation supports the idea that associative models better match human behaviour [9], [11]. An approach based on a Bayesian framework has been proposed by [12] which is moreover able to integrate social cues in learning and was shown to outperform a simple associative model on a dataset of parent-children interactions. This problem has also been studied using the so-called Neural Modeling Fields [13] showing the capacity of this generic and more biologically plausible approach to solve the problem.

All these models take symbolic data as input, both for the words and for the referents and focus on learning their correct associations. Other models rely on non-symbolic perception, both for the referents and/or for the words, and thus are closer to implementation on robots with real sensors. For example, the talking heads experiment [14] uses a camera to observe coloured shapes and shows how to create a grounded meaning of symbols shared between robots. Though not originally designed for cross-situational learning, it has been shown that it can be extended to this case [15] and is able to learn word meanings, both for objects and their features (size, position, color). The CELL model [16] is also able to learn word-meaning association through multimodal learning. Working on two continuous modalities, it first segments each modality independently and then learns associations between these segments by analysing their co-occurences and mutual information on a short-term memory in order to produce word-meaning clusters. The model of Yu and Ballard [17] learns word-meaning in a multi-modal setup using continuous speech and observation of actions performed by humans. Although the spoken language is constrained and several social cues are used, the model is able to learn objects' names, but also verb meanings. Mangin and Oudeyer [18] also proposed a multi-modal learning approach that is able to discover elementary gestures performed by a human and their names in continuous audio and motion streams.

Several algorithmic approaches have been proposed for the problem. These approaches may be either very specific to the problem [6], [10], or take advantage of standard approaches such as the Bayesian approach [12], the Expectation-Maximisation algorithm [17] or the Neural Modelling Fields [13]. Topic discovery algorithm is also an interesting approach for this problem as it can discover the underlying element (the object or feature) that will generate the perception and the associated word. Among the existing algorithms, Non Negative Matrix Factorization has been used in [18] and Latent Dirichlet Allocation in [19].

III. Proposed approach

Our approach for word-meaning learning uses symbolic information for the language and continuous data for the visual perception. These channels are both noisy and the language part is ambiguous as it contains many words not related to the object identities or features. We use two complementary algorithms: one based on statistical analysis of word-object occurrences to filter noise in the symbolic channel and the other based on Non Negative Matrix factorization [20] to discover the word-meaning associations from the multi-modal data gathered across different situations.

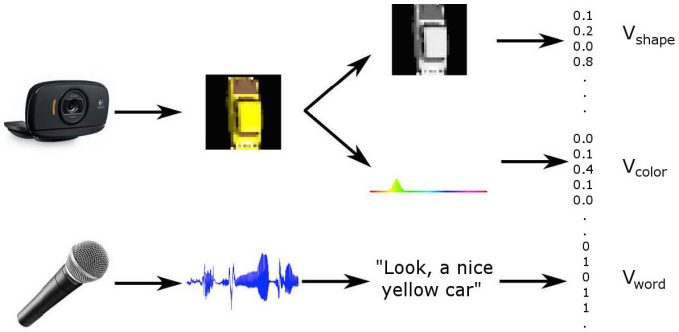


Fig. 1. Data representation used in our experiments

A. Data representation

We have two modalities as input to our model (Figure 1). The first one is a noisy continuous modality that typically represents different features of the object of interest obtained through computer vision. These features are currently constructed to represent color and shape of the object (see section IV-A), but they could also be the results of a more generic feature computation algorithm, for example learned through deep-learning. The features are encoded as vectors of constant size, and they do not present ambiguity in the sense that they only represent one object of interest that is designated by the human partner. However, the noise in the features could sometimes prevent the correct recognition of an object.

The second channel is a noisy symbolic channel and represents the word occurrences in a sentence. It is typically taken from a real speech recognition system and therefore often contains erroneous word recognitions, besides the real words in the sentence that are not related to the object of interest (such as verbs and articles). The word occurrences are coded as a binary vector of the size of the dictionary of all known words. The dictionary is created incrementally, starting from an empty dictionary and adding each new word encountered in sentences at the end.

The system is presented with a set of observations of objects and sentences pronounced by a human partner. Each situation is encoded as a vector concatenating the two modalities.

B. Symbolic channel filtering

The first part of our model is in charge of filtering noise in the symbolic channel. It relies on statistics on the word occurrences processed through the Term Frequency-Inverse Document Frequency (TF-IDF) approach [21] popular in text processing.

We first group the observations according to their similarities in the non symbolic channel. Intuitively, the goal is for the system to put together all the observations that share a common feature in order to be able to analyse the associated word statistics. The clustering is performed by a simple incremental clustering that puts each observation in the same cluster as a previous observation if its distance is smaller than a threshold (we used 0.8 in our experiments), or creates a new cluster otherwise. We use the χ^2 distance which is well adapted for histogram features :

$$\chi^2(x,y) = \sum_{k=1}^{d} (x_k - y_k)^2/(x_k + y_k)$$

The *term-frequency* of the word i associated with each cluster j is then computed :

$$tf_{ij} = n_{ij}/n_j$$

where n_j is the total number of words observed in samples from cluster j, and n_{ij} is the number of occurrence of word i observed in samples from cluster j. This value is high for words that occur often with a given object. The words with tf below or equal a threshold (we chose the second highest tf value for each cluster) are considered as noise and removed from the observations (their entry is put to 0).

The *inverse document frequency* of the remaining words i is then computed:

$$idf_i = log[N/(1+N_i)]$$

where N is the number of clusters, and N_i is the number of clusters where word i appears at least once. This measure is high for words that appear in very few clusters and low when they appear in many clusters. The words with idf above a threshold are common words (such as articles) and removed from the observations. The words with idf below a threshold are also removed, making the assumption that each word will eventually be associated to several different objects (e.g. two objects will have the same color). Assuming a mean repartition of 2 keywords for each object, the mean idf value should be $log[N/(1+\sqrt{N})]$. We therefore use $log[N/(1+\sqrt{N}\pm\epsilon)]$ for the high and low threshold ($\epsilon = 2$ in our experiments).

The remaining words after these two steps should contain little noise and represent the words for which we have enough cross-situational data to learn their meaning.

C. Learning meaning through Non Negative Matrix Factorization

Using the samples filtered in the previous step, we use Non-negative Matrix Factorization(NMF) [20] in order to discover reference vectors that explain data efficiently as sum of these reference vectors with positive weights. This method has been shown to be able to discover part-based object representation [22], which is close to our problem as we want to discover part of the feature vector (meaning) associated with part of the dictionary (word). More precisely, NMF will find matrices W and H so that :

$$\begin{aligned} V_{m\times n} &= W_{m\times k} H_{k\times n} \\ \begin{bmatrix} V_{shape} \\ V_{color} \\ V_{word} \end{bmatrix}_{m\times n} &= \begin{bmatrix} W_{shape} \\ W_{color} \\ W_{word} \end{bmatrix}_{m\times k} [H_1, H_2, ..., H_n]_{k\times n} \end{aligned} \quad (1)$$

where V is the matrix containing all the observations in columns, W and H are the matrices computed by NMF, W containing k reference elements and H being the weights that make it possible to reconstruct the observations from the reference elements. As we have symbolic labels in the language modality (words), and we want to have a meaning for each of these words, we set the parameter k to the number of words that remain after the first filtering step.

The W and H matrices are then computed using the algorithm based on multiplicative update proposed by Lee and Seung [22] and used in [18] for similar problems. This method converges to a local minima, so the initialisation is important. In our case, we want to discover word-meaning association, so we initialise the W_{word} matrix to the identity so that it favours solutions with one word for each reference element. We initialise W_{shape} and W_{color} to random values.

The result of this algorithm is a set of k vectors associating word activations and their meanings in the feature representation part. Our approach therefore provides an explicit meaning for all the words retained after the TF-IDF filtering.

D. Incremental learning

In order to perform incremental learning after each new observation, we add the new observation to the matrix V, adjusting the size of the matrix if new words appeared or were removed by the TF-IDF filtering. We initialize matrices W and H with the results of the previous time step, augmented with a random column if a new word has passed the filtering process. We perform NMF updates until the magnitude of the update falls below a threshold ($10^{-6}mn$, where m, n are from Equation 1) or a maximum number of iteration (200) is reached.

IV. Experimental results

We now present the experimental setup, a particular speech and image processing used to create learning samples and the results on word-meaning association learning.

A. Experimental setup

The experiment is conducted with a camera installed over a table, facing down to capture the image of objects and a microphone is used for acquisition of participants' vocal sentences, which will be converted into text format. 24 objects (Figure2) can be put on the table, one at a time, while a human teacher is describing them. All 24 objects can be grouped by color as "blue (*bleu* in French)", "green (*vert*)", "red (*rouge*)", "yellow (*jaune*)", and categorized by shape as "cup



Fig. 2. The 24 objects used for the experiments.

(*tasse*)", "ring (*anneau*)", "lego", "apple (*pomme*)", "compass (*boussole*)", "car (*voiture*)", "book (*livre*)". Our experiments aimed at learning these 11 word meanings.

For image processing, we used a simple approach under controlled conditions to perform the first experiments validating our overall approach. The OpenCV[1] library was used to segment the object from the background using a simple threshold on pixel intensity (the background being black). In order to obtain comparable shape information of objects, reference *angular position* and *size* are defined by fitting objects with their smallest rectangular bounding box, rotating it to be parallel to axis before resizing it to a size of 30×30 pixels. The shape feature vector is constructed from this gray-scale converted image, taking lines one after another to form a 900 elements feature vector. The color feature is constructed from the Hue value in the HSV (Hue Saturation Value) color space to construct a 80 bins histogram that is additionally smoothed using a Gaussian Filter. The complete visual feature vector is therefore of size 980, representing both visual features and can be easily interpreted visually (see figure 3), which simplifies the qualitative analysis of the results.

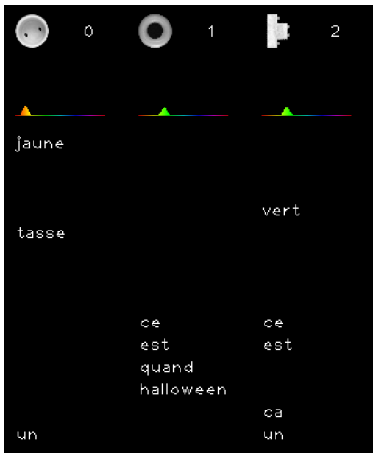


Fig. 3. Examples of correct, corrupted and half-correct samples

[1]http://www.opencv.org

The speech-text conversion relies on Google speech-api[2] which is used to return the recognized sentence in text format. The recognition is applied to whole sentences of a teacher for the description of an object. From this text, we build a binary word histogram as described in section III-A.

We recorded from three different teachers a total of 77 samples, manually categorized as *correct* (i.e. containing both correct labels of color and shape along with other unrelated words), *half-correct* (i.e. containing only one of color or shape label) and *corrupted* (i.e. containing no correct label) so as to analyse our method's robustness (see Figure 3). These samples are replayed in random order to get performance statistics. Figure 3 illustrates these samples. "Une tasse jaune" correctly describes a yellow cup. When describing a green lego, the speaker says: "Ha, c'est un lego vert", however the word "lego" fails to be recognized and leads to a half-correct sample on the right. A corrupted sample appears when none of the correct words are recognized, for example when the original sentence "c'est un anneau vert" is misunderstood as "c'est quand halloween". Among all recorded samples, there are 70.13% correct samples, 11.69% half-correct samples and 18.18% corrupted samples. Overall, the 11 words to learn represent only 17.74% of the total number of words present in the samples.

B. Learning with perfect symbolic labels

In this first experiment we use only perfect symbolic labels (therefore using no symbolic label filtering) to validate that NMF can learn label meanings.

The training set include samples of nine of the objects in Figure 4, covering three color and three shape symbols. From this simple case, we manually constructed a reference dictionary of symbols whose feature descriptions, regarding either shape or color, are averaged values of symbol-related samples from recorded data (see Figure 4). As the number of real word to learn is known to be 6 in this setup, we choose $k = 6$ for the NMF algorithm.

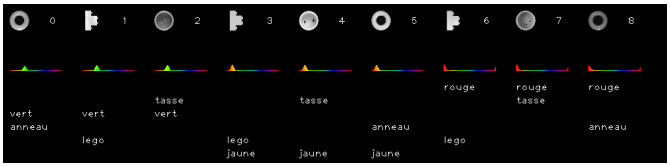
As shown in Figure 4, the proposed NMF method is able to extract symbolic labels from samples and the learned dictionary approximates very closely the reference dictionary, with only limited noise in the shape description part.

C. Learning with noisy word recognition

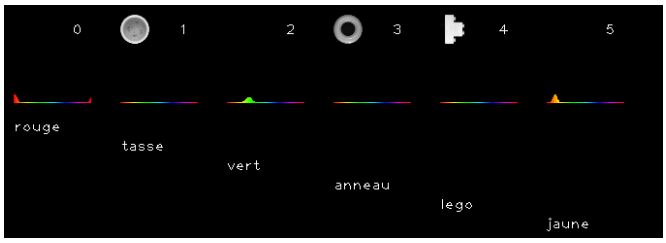
We now perform experiments with all 77 samples, using directly the speech recognition results with our symbolic label filtering method. We performed 10 experiments by processing all the samples in random order and report mean and variance of values computed for these 10 experiments.

Figure 5a plots the evolution of the total number of different words encountered in the samples (in black), the number of selected words by our filtering scheme (in red) and the real number of correct words in these selected words (in blue). We can see that our approach selects an approximately correct number of words during the whole experiment, and converge

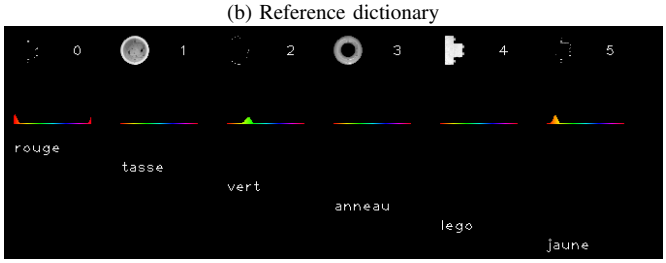
[2]https://github.com/gillesdemey/google-speech-v2



(a) Input training samples



(b) Reference dictionary



(c) Dictionary learned with NMF

Fig. 4. Learning with manually created perfect symbolic labels. The learned dictionary using NMF is very close to the reference dictionary created by averaging features for the corresponding words.

to the correct total number of keywords after approximately 50 samples.

We also defined two metrics for the evaluation of the quality of word filtering and word-meaning dictionary. For the word filtering part, we compare the set of filtered words F with the set of reference words R using the equation:

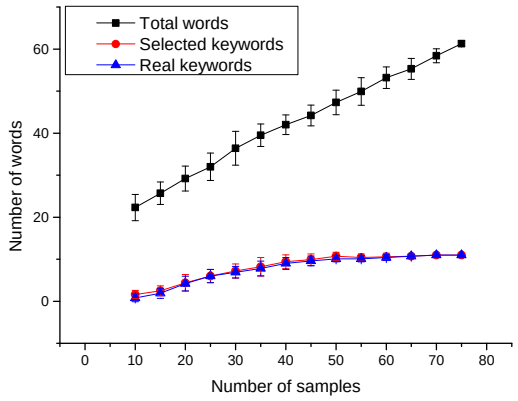
$$s_{word} = 100 \times \frac{Card(F \cap R)^2}{Card(F) \times Card(R)} \tag{2}$$

where $Card(X)$ is the number of elements of X. This score is maximal when $F = R$ and decreases when the set of filtered words lacks some reference elements or when it contains additional erroneous words.

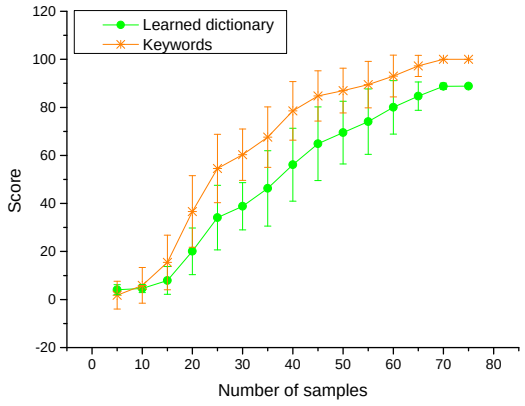
We compare the learned dictionary with the reference dictionary (see section IV-B) using the following formula:

$$s_{dict} = 100 \times \frac{\left(\sum_{i \in R} \sum_{j \in F} \delta(i,j) \cdot e^{-\chi^2(r_i, f_j)} \right) \cdot Card(F \cap R)}{Card(F) \times Card(R)} \tag{3}$$

where $\delta(i,j)$ is the Dirac function that equals 1 when the most activated word of the learned dictionary entry j is the same as in the reference word i, 0 otherwise. $\chi^2(r_i, f_j)$ is the χ^2 distance between the visual feature part of learned entry j and reference entry i. This measure is maximum when the learned dictionary is equal to the reference dictionary and decreases when the selected words are different from the reference or



(a) Performance of keywords's filtering



(b) Quality of keywords and dictionary learning

Fig. 5. Learning results with full sample set, using word filtering and NMF: a) word filtering successfully detects the 11 keywords. b) NMF produces high quality dictionaries after 70 samples.

when the definition of the selected words differ from their reference definition.

Figure 5b shows the mean and variance of these values with incremental amount of training samples. We can see that the word filtering improves its performance other time and reach a maximum score after 70 samples in all cases. The dictionary quality follows very closely the word filtering quality, showing that the dictionary learning using NMF is efficient and that the overall quality mainly depends on the word filtering. The difference with 100 is mostly due to remaining noise in the shape description. The resulting learned word-meaning dictionaries are not shown due to limited space but are qualitatively very similar to the one in Figure 4c.

V. Discussion and Conclusion

We proposed an algorithmic approach to learn word-meaning associations in a cross-situational setup with noisy and ambiguous input taken from vision and speech recognition. This approach was shown to be robust in preliminary real interactive experiments, dealing efficiently with the errors

and unrelated words produced by speech-recognition (correct words represents only 17.74% of total words) and the variations occurring in visual data.

The proposed approach, while not aiming at modelling human behaviour, is compatible with the hypothesis that humans use associative methods for cross-situational learning [9], [11]. Like these models, it uses statistics on word occurrences and their associations with referents. However, it goes beyond such models by being able to process non-symbolic referents and using statistical correlations discovered by NMF to define the word meaning instead of associating words with predefined symbolic referents. It shares this capability with other models [16], [17], but relies on a separate speech recognition technology, benefiting from its performance, while solving their shortcomings by filtering erroneous recognitions.

Our approach extends [18] (who also used NMF in related tasks), by being able to learn with very noisy input thanks to the TF-IDF words filtering. It also shows that NMF can learn word-meanings that appear in one modality only (e.g., red is correlated with color only) while [18] assumed that the learned concepts always have manifestation in all modalities. Our model also has the advantage of providing an explicit representation of each word meaning in the elements learned by NMF. While this is not compulsory as concept learning can be implicit and observed from robot or human behaviour (as in [18]), it makes it possible to define a clear quality measure as used in section IV. However, for better comparison with other approaches, we plan to assess the performances of the learned dictionaries by measuring object and word recognition rates on separate test databases.

Our model also makes it possible to learn the meaning of different types of words relating to the object identity (nouns) or its features (adjectives). While this remains quite limited with respect to the complexity of natural language, it is interesting to be able to learn these two different kinds of words without specific processing. Going further in complexity could take advantage of the language structure itself [5] to guide the word-meaning associations.

In future work, we also plan to extend our approach to ambiguity in the visual modality by using more complex image processing that would be applicable on autonomous robots in indoor environments and by having several objects shown at the same time. This is simple for color, which is an additive feature when several objects are presented and thus can be processed directly by NMF, but it will require using other features for shape as the current one is not additive. We also plan to extend our approach to deal with homonyms, both for the language part and for the visual part, where an object can present different visual appearances depending on the observation point of view.

ACKNOWLEDGMENT

The authors would like to thank Fabio Pardo and Olivier Mangin for their help in implementing the reported experiments. This work was supported by the China Scholarship Council.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, pp. 335–346, 1990.

[2] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558 – 1568, 2008.

[3] W. V. O. Quine, *Word & Object*. The Mit Press, 1960.

[4] M. Hirotani, M. Stets, T. Striano, and A. D. Friederici, "Joint attention helps infants learn new words: event-related potential evidence," *Neuroreport*, vol. 20, pp. 600–605, 2009.

[5] L. Gleitman, "The Structural Sources of Verb Meanings," *Language Acquisition*, vol. 1, no. 1, pp. 3–55, 1990. [Online]. Available: http://dx.doi.org/10.2307/20011341

[6] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, no. 12, pp. 39 – 91, 1996, compositional Language Acquisition.

[7] C. Yu, L. B. Smith, C. Yu, and L. B. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychological Science*, pp. 414–420, 2007.

[8] G. Kachergis and C. Yu, "More naturalistic cross-situational word learning," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 2013.

[9] G. Kachergis, C. Yu, and R. Shiffrin, "Cross-situational word learning is better modeled by associations than hypotheses," in *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, Nov 2012, pp. 1–6.

[10] T. N. Medina, J. Snedeker, J. C. Trueswell, and L. R. Gleitman, "How words can and cannot be learned by observation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9014–9019, 2011. [Online]. Available: http://www.pnas.org/content/108/22/9014.abstract

[11] G. Kachergis and C. Yu, "Continuous measure of word learning supports associative model," in *Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on*, Oct 2014, pp. 20–25.

[12] N. Goodman, J. B. Tenenbaum, and M. J. Black, "A bayesian framework for cross-situational word-learning," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Curran Associates, Inc., 2008, pp. 457–464.

[13] J. F. Fontanari, V. Tikhanoff, A. Cangelosi, R. Ilin, and L. I. Perlovsky, "Cross-situational learning of objectword mapping using neural modeling fields," *Neural Networks*, vol. 22, no. 56, pp. 579 – 585, 2009.

[14] L. Steels, "Evolving grounded communication for robots," *Trends in Cognitive Sciences*, vol. 7, no. 7, pp. 308 – 312, 2003.

[15] J. De Beule, B. De Vylder, and T. Belpaeme, "A cross-situational learning algorithm for damping homonymy in the guessing game," in *Artificial Life X : Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, L. M. Rocha, L. S. Yaeger, M. A. Bedau, D. Floreano, R. L. Goldstone, and A. Vespignani, Eds., International Society for Artificial Life. The MIT Press (Bradford Books), Aug. 2006, pp. 466–472.

[16] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive science*, vol. 26, pp. 113–146, 2002.

[17] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Trans. Appl. Percept.*, vol. 1, no. 1, pp. 57–80, Jul. 2004. [Online]. Available: http://doi.acm.org/10.1145/1008722.1008727

[18] O. Mangin and P.-y. Oudeyer, "Learning semantic components from sub-symbolic multi-modal perception," in *Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*, August 2013.

[19] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using lda," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 3943–3948.

[20] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing ...*, pp. 556–562, 2001. [Online]. Available: http://papers.nips.cc/paper/1861-alg

[21] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988. [Online]. Available: http://dx.doi.org/10.1016/0306-4573(88)90021-0

[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, pp. 788–791, 1999.