

The Learning of Adjectives and Nouns from Affordance and Appearance Features

Onur Yürüten¹, Erol Şahin² and Sinan Kalkan²

June 21, 2013

Abstract

We study how a robot can link concepts represented by adjectives and nouns in language with its own sensorimotor interactions. Specifically, an iCub humanoid robot interacts with a group of objects using a repertoire of manipulation behaviors. The objects are labeled using a set of adjectives and nouns. The effects induced on the objects are labeled as affordances, and classifiers are learned to predict the affordances from the appearance of an object. We evaluated three different models for learning adjectives and nouns using features obtained from the appearance and affordances of an object, through cross-validated training as well as through testing on novel objects. The results indicate that shape-related adjectives are best learned using features related to affordances, whereas nouns are best learned using appearance features. Analysis of the feature relevancy shows that affordance features are more relevant for adjectives and appearance features for nouns. We have shown that adjective predictions can be used to solve the odd-one-out task on a number of examples. Finally, we linked our results with studies from Psychology, Neuroscience and Linguistics that point to the differences between the development and representation of adjectives and nouns in humans.

Keywords: affordances, nouns, adjectives

1 Introduction

Seamless communication with humans is an ambitious challenge for robots that requires linking linguistic categories (such as nouns and adjectives) to the sensorimotor interactions of the robot. The gap between the discrete symbols of language, such as nouns and adjectives, and the continuous

¹ Department of Computer and Communication Sciences,
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

²KOVAN Research Lab, Dept. of Computer Engineering, Middle East Technical University, Ankara, Turkey

Corresponding Author: Onur Yürüten EPFL EDOC-IC, INR 018 (Batiment INR), Station 14 CH-1015, Lausanne
Email: onur.yuruten@epfl.ch

and high-dimensional stream of sensorimotor data was pointed out by Harnad (Harnad, 1990) as the *symbol grounding problem*. Experimental data (Borghi, 2007; Cangelosi and Harnad, 2001; Fischer and Zwaan, 2008; Gallese and Lakoff, 2005) has supported the view that language is grounded in the sensorimotor experiences of the organism (Cangelosi and Riga, 2006; Cangelosi et al., 2010; Steels, 2003; Glenberg and Kaschak, 2002; Cangelosi, 2010) and that understanding words requires the very same neural circuitry that is active during the sensorimotor interactions of the organism related to the meaning of the word (Glenberg et al., 2008; Zwaan and Taylor, 2006).

The question that we address in this paper can be simply put forward as: *How can a robot ground nouns and adjectives into its own sensory-motor interactions?* That is, when we command the robot to grasp the *tall cup*, how can it map the symbols *tall* and *cup* into its perceptual view of the world.

2 Robots, Language and Affordances

The issue of what exactly nouns, such as *strawberry*, represent is subject to debate. One view argues that nouns are categories formed by the visual appearance of objects, whereas another argues for categories based on the function of objects (Borghi et al., 2002). However, there is evidence that humans use both mechanisms which operate in an interconnected manner based on context or goal (Helbig et al., 2006; Borghi et al., 2002). The hypothesis that there might be distinct mechanisms for categorizing objects is further supported by studies in neuroscience and psychology which suggest that objects are processed through two different pathways (Goodale and Milner, 1992), one involving the Object Recognition (OR) system categorizing an object based on its visual appearance, the other involving the affordances that the object offers to the organism.

The categorization of objects based on their visual appearances is a well-studied and hot topic in computer vision. However, such approaches often fail to capture the essence of a noun such as *chair* which may appear in very different forms (see Figure 1 for some examples) and is beyond the focus of our study. Since objects provide certain functionalities to us in our daily lives, objects can also be categorized by other aspects that are consistently related to their function (Borghi et al., 2002). This function-based categorization can be used along with the appearance-based categorization of objects.

There have been many attempts at linking nouns to the sensorimotor experiences of robots in the robotics community. For example, Yu and Ballard (2004) proposed a system mapping words in speech to co-occurring features in images using a generative correspondence model. Carbonetto and de Freitas (2003) presented a system that splits a given image into regions and finds a proper mapping between regions and nouns inside the given dictionary using a probabilistic translation mode similar to a machine translation problem. From a different perspective, Cangelosi et al. (2010) suggested an interactive approach to learning lexical semantics by demonstrating how an agent can use heuristics to learn simple shapes which are presented

by a tutor with unrestricted speech. Their method matches perceptual changes in a robot’s sensors with the spoken words and uses a k-nearest neighbors algorithm in order to learn the names of shapes. In similar studies, Morse et al. (2011); Cangelosi and Parisi (2004) use neural networks to link words with behaviors of robots and with extracted visual features.

For learning adjectives from the sensorimotor interactions of the robot, many attempts have recently been made. McMahon et al. (2012) developed a method for learning haptic adjectives from interactions whereas Petrosino and Gold (2010), Dindo and Zambuto (2010), and Chella et al. (2009) studied learning color, size and distance related adjectives based on visual features. Similar studies (Chauhan and Lopes, 2011; Haazebroek et al., 2011; Sugita et al., 2011; Glenberg and Gallese, 2011; Morse et al., 2011; Gold et al., 2009) proposed methods for learning object categories; however, systematic evaluation of nouns and adjectives based on appearance and affordances has not been performed previously.



Figure 1: Objects with various differences, yet all called *chair*.

2.1 Affordances and Language

The notion of affordances was introduced by Gibson to explain how inherent “values” and “meanings” of things in the environment can be directly perceived and how this information can be linked to the action possibilities offered to the organism by the environment (Gibson, 1986).

The link between affordances and language comprehension has already been pointed out in Psychology. The indexical hypothesis (Glenberg and Robertson, 2000) claims that words are linked to entities or objects in the real world, or to representations such as pictures or perceptual symbols (Barsalou, 1999). For example, the word *can* is linked to its referent, a can, or to an analogical representation of a can. Therefore, words that refer to objects would initially activate perceptual information corresponding to such objects from our previous experiences.

Since perceptual and motor processes are tightly linked, one expects that words should also activate motor information. In fact, depending on their perceptual features, objects can activate affordances (Bub et al.,

2008; Jax and Buxbaum, 2010; Pellicano et al., 2010). For instance, different kinds of cans may afford different actions: some can be filled-in, some can be used as steps. This suggests that affordances of objects can be linked to words representing these objects.

Inspired by the notion of affordances, Montesano et al. (2008a, 2009) proposed a probabilistic model to encode the relations between objects, actions and observed effects. Through imitation games, they show that such a representation can be used to learn what objects afford and hence provide an implicit way of categorizing objects. Their work is further extended in Salvi et al. (2012) towards associating the meaning of words with affordance information. Sun et al. (2010) studied the categorization of objects in terms of probabilistic relations between objects and affordances in a scenario where traversability of different types of floors and objects was learned.

Although the relationship between language and affordances has been pointed out by many (such as Gibson (2000)), the issues of how such a link can be created in robots has not yet been fully tackled.

In a similar way, for categorizing objects, one can consider in what kind of activities or sequences of actions objects are utilized. Aksoy et al. (2010) use how spatial relations between image segments change in time to cluster action sequences, and categorize objects based on the action sequences in which they are used. Likewise, Wu and Aghajan (2009) approach the object recognition problem by modeling the prior knowledge of the relationship between the users activity and objects. They approached the problem with an explicit modeling of the environment, leading to a computationally expensive methodology. The key points in these approaches lie in selecting or designing state-of-the-art features and descriptors, and then linking them with functionalities through symbolic descriptions that represent actions.

2.2 The Current Study

Learning noun and adjective categories has been mostly studied as a mapping of the visual appearance of objects to linguistic categories. The few exceptions focused only on categorizing objects based on the kinds of tasks or activities they are used in (see, e.g., Aksoy et al. (2010); Wu and Aghajan (2009)). Another exception is our previous study (Dağ et al., 2010; Atıl et al., 2010) in which we categorized objects based on their affordances. The similarities and differences between learning nouns and adjectives from the appearance and affordances of objects have not been investigated in the literature.

In this study, we are concerned with linking adjectives and nouns with the sensorimotor experiences of the humanoid robot iCub. Towards this end, we use the notion of affordances by Gibson (1986) as formalized by Şahin et al. (2007), and we provide a systematic evaluation of nouns and adjectives based on appearance and affordances.

In this article we focus only on concrete nouns (nouns that can be directly linked to physical entities in the world). The results and the conclusions drawn in the article do not apply to abstract nouns (such as joy, hatred). For the distinction between concrete and abstract nouns we refer

to Borghi et al. (2011), Crutch and Warrington (2005), and Pexman et al. (2007). Moreover, there is a syntactical difference between nouns and adjectives - see, e.g., Baker (2003). In this article, we only focus on their differences in terms of learning.

The current article is an extension of our previous study (Yuruten et al., 2012) in the following aspects: (i) the computational analyses are extended in terms of the models used as well as their extensive comparison with respect to using appearances or affordances, (ii) the models are tested on a publicly available 3D object database (the KIT 3D object database - Kasper et al. (2012)), (iii) the models are applied to the “Odd-one Out” task, and (iv) the distinction between adjectives and nouns in their preference for affordances or appearances is discussed in depth in relation to findings from Psychology and Neuroscience, and Linguistics.

3 Affordance Formalization

In Şahin et al. (2007), we argued that each interaction episode of an agent with its environment would create an affordance relation instance between three sets of information as,

$$(entity, behavior, effect), \quad (1)$$

where *entity* denotes the perceived information from the environment obtained and the robot itself¹. *Behavior* denotes the means of interaction for the robot, and finally, *effect* is the change in the environment generated when the behavior is executed (for similar formalizations, see, e.g., Montesano et al. (2008b); Krüger et al. (2011)). For instance, a robot applying its lift-with-right-hand behavior on a blue-can to generate a lifted effect can be represented with a relation instance as:

$$(blue-can, lift-with-right-hand, lifted), \quad (2)$$

where the terms *blue-can*, *lift-with-right-hand*, and *lifted* are merely placeholders for the corresponding perceptual and proprioceptive representations. However, a single relation instance provides little predictive ability over future experiments, such as whether the application of the same behavior on a red-can or a blue-desk will generate the same effect or not. Only after interacting with other objects, such as a green-can, can one join the relation instances together as:

$$\left(\begin{array}{l} blue-can \\ green-can \end{array} \right), lift, lifted). \quad (3)$$

These types of instances can be abstracted by a mechanism to produce an abstract entity equivalence class like:

$$(< can >, lift-with-right-hand, lifted), \quad (4)$$

where $<can>$ represents the derived invariants of the entity equivalence class. In this example, $<can>$ means “cans of any color” that can be lifted upon the application of lift-with-right-hand behavior. These types

of abstractions create a general affordance relationship, which allows the robot to predict the effect of the lift-with-right-hand behavior applied to an unseen object like a *red-can*. Such a skill facilitates great flexibility in a robot. For example, when needed, the robot can search and find objects that would provide support for a desired affordance. We argue that the creation of equivalence classes covering the three components of the relation provides a mechanism for creating abstract categories that can be linked to concepts represented by nouns and adjectives.

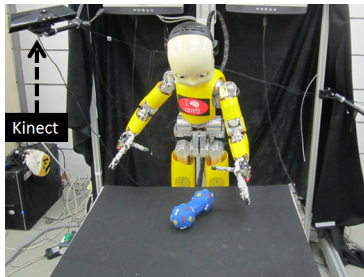


Figure 2: iCub interacting with an object.

4 Experimental Framework

We used the iCub robot platform (Metta et al., 2008), a 53 DoF humanoid in the form and size of a 3.5-year-old child, in our study. In order to perceive the environment, a fixated Kinect range camera was used as shown in Figure 2. The camera captured the point cloud images of the table placed in front of the robot with a resolution of 640×480 at 30fps.

4.1 Behaviors

The robot interacted with the objects using a repertoire of six manipulation behaviors: *push-left*, *push-right*, *push-forward*, *pull*, *top-grasp* and *side-grasp*. The *push-** and *pull* behaviors pushes or pulls back the object in the stated direction. Both *top-grasp* and *side-grasp* behaviors are used to grasp the object. In the former, the robot approaches the object from top, whereas in the latter one the robot approaches the object from the side. The behaviors are similar to the ones used by Bergquist et al. (2009), and Metta and Fitzpatrick (2003).

4.2 Objects

The objects (35 in total) are labeled with three adjectives picked from $\mathcal{D} = \{edgy \times round, short \times tall, thick \times thin\}$ and one noun from picked $\mathcal{N} = \{ball, cylinder, box, cup\}$ as shown in Figures 3 and 4.

The co-occurrences between nouns and adjectives as well as within adjectives are shown in Table 1 and Table 2 respectively. It can be seen

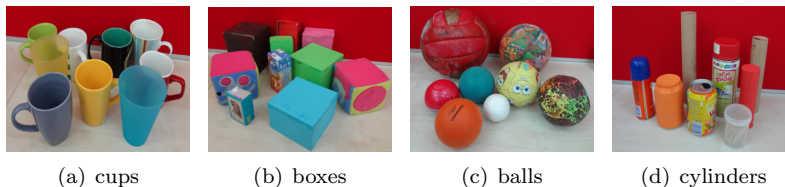


Figure 3: Objects grouped by the nouns.

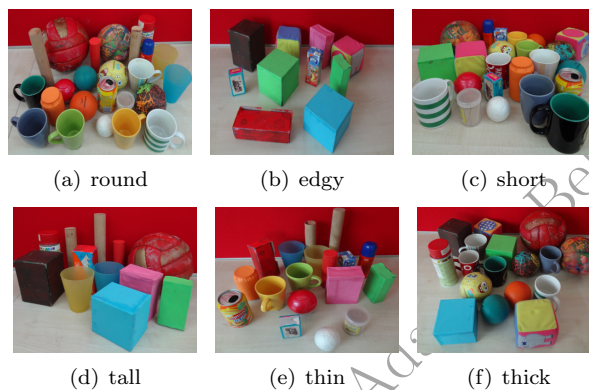


Figure 4: Objects grouped by the adjectives. Note that cups that have varying width (from *thin* to *thick*) are labelled as *thin*.

that no adjective-adjective or noun-adjective pairs have a 100% correlation. This means that, for instance, if an object is known to be *round*, neither will it imply the object to be a *ball*, nor will it imply it to be *thin* or *thick*.

4.3 Perceptual features

Objects placed on the table are segmented from the rest of the scene by filtering out the planar table-top. This is achieved by filtering out the 3D points that fall outside of the table. The point cloud of the object is then processed to extract its perceptual representation which consists of:

Table 1: Co-occurrences between adjectives and nouns in our dataset.

Noun	Edgy	Round	Short	Tall	Thin	Thick	Total
Ball	0	8	7	1	6	2	8
Cup	1	8	7	2	4	5	9
Box	10	0	5	5	5	5	10
Cylinder	0	8	4	4	4	4	8

Table 2: Co-occurrences between adjectives in our dataset.

Adjective	Edgy	Round	Short	Tall	Thick	Thin
Edgy	x	0	9	2	6	5
Round	0	x	14	10	10	14
Short	9	14	x	0	11	12
Tall	2	10	0	x	5	7
Thick	6	10	11	5	x	0
Thin	5	14	12	7	0	x

- *Surface features*: surface normals (azimuth and zenith angles - two 20-bin histograms), principal curvatures (20-bin histogram), and shape index (20-bin histogram) using the methods provided by the Point Cloud Library (Rusu and Cousins, 2011). These features provide information about the 3D shape of the object. Principal curvatures K_1, K_2 at a point give information about how the surface normal changes along two orthogonal directions at the point. Shape index is effectively a combination of the two principal curvatures $(\frac{K_1+K_2}{K_1-K_2})$, and a compact measure for describing the surface type at a point.
- *Spatial features*: bounding box, center, orientation, and dimensions (along x, y, z). The orientation of the object is determined by finding the principal axes of the point cloud. The bounding box is then determined by finding the farthest points along the principal axes. The center is simply the average of the points in the point cloud. Finally, elongations along the x, y and z axes are taken as the dimensions of the object.
- *Object Presence*: a binary feature indicating the presence/absence of an object on the table. This is calculated by checking whether there are any 3D points on the table or not.

Hence, the object is represented by a perceptual vector of size 88, denoted by \mathcal{V}_E .

5 Methods

The robot made 413 interactions with objects placed at random orientations and positions on the table, using its behavioral repertoire. Each interaction episode is encoded as an affordance relation between an object $o_j \in \mathcal{O}$, a behavior $b_i \in \mathcal{B}$ and an effect f as

$$(e_{o_i}, b_j, f_{o_i}^{b_j}),$$

where $e_{o_i} \in \mathcal{V}_E$ is the initial perceptual representation of the object o_i , and $b_j \in \mathcal{B}$ is a behavior². $f_{o_i}^{b_j}$ denotes the effect label provided by the user from the set $\mathcal{F} = \{no\ effect, moved\ left, moved\ right, moved\ forward,$

Table 3: Notations used in the article.

Notation	Meaning
o	An object.
\mathcal{O}	The set of objects used in the experiments.
e_o	Perceptual feature of object o .
b	A behavior the robot is equipped with.
\mathcal{B}	The set of behaviors the robot is equipped with.
f	An effect that can be generated on an object.
\mathcal{F}	The set of effects that the robot can generate.
\mathcal{V}_E	Perceptual feature vector (E for entity).
\mathcal{V}_A	Affordance feature vector (A for affordance).
\mathcal{V}_C	Combined vector (C for combined).
M	A mapping from one space to another.
M_X^Y	A mapping from V_X to Y , where $X \in \{E, A, C\}$ and $Y \in \{\mathcal{N}, \mathcal{D}\}$.
\mathcal{N}	The set of nouns.
\mathcal{D}	The set of adjectives.

Table 4: Set of behaviors, nouns, adjectives and effect labels.

Behaviors	Nouns	Adjectives	Effect Labels
<i>push-left</i>	<i>ball</i>	<i>edgy vs. round</i>	<i>moved right</i>
<i>push-right</i>	<i>box</i>	<i>short vs. tall</i>	<i>moved left</i>
<i>push-forward</i>	<i>cup</i>	<i>thick vs. thin</i>	<i>moved forward</i>
<i>pull</i>	<i>cylinder</i>		<i>pulled</i>
<i>top-grasp</i>			<i>knocked</i>
<i>side-grasp</i>			<i>no effect</i>
			<i>grasped</i>
			<i>disappeared</i>

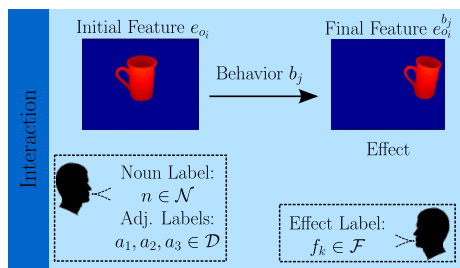


Figure 5: In an interaction the robot observes an object, applies a behavior and induces an effect. During the interaction a human provides an effect label, a noun label, and three adjective labels.

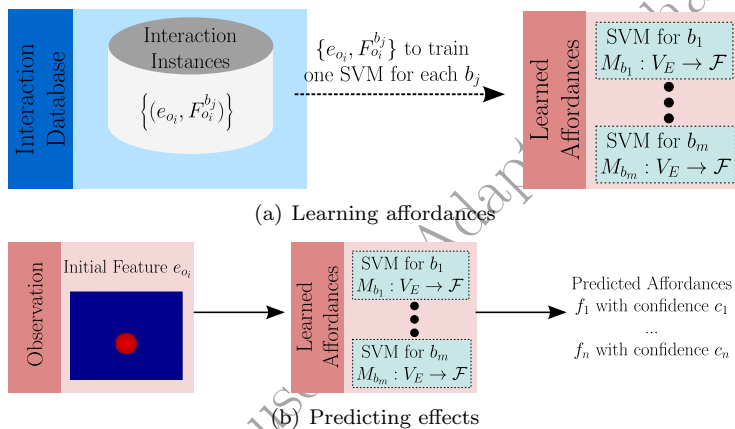


Figure 6: Learning and predicting affordances. (a) For each behavior, an SVM is trained to link the perceptual features of the object to the effect labels. (b) The trained SVMs can be used to predict the affordances of an object.

pulled, grasped, knocked, disappeared. These labels correspond to the set of affordances that the robot has within the experimental setup.

Note that the behaviors induce different effects on the objects based on their type as well as on their position and orientation. For example, the *push*-* (any of the push behaviors) behaviors would make a *ball* placed anywhere on the table roll away and *disappear(ed)*. However the same behaviors would make *boxes disappear* when they are placed on the edges of table and pushed in the right direction, but would cause the *boxes* to be *moved*-* if they are placed at the center of the table.

5.1 Learning Affordances

For each behavior, we train a classifier that maps the initial perceptual view of the object to the effect label using the data collected from the interactions. Specifically, we first analyze the data using the ReliefF al-

Table 5: Sample confidences of obtaining each effect label from each behavior on an object. The matrix is transformed into a vector called the affordance vector.

Behaviors vs. Effects	<i>push-right</i>	<i>push-left</i>	<i>push-forward</i>	<i>pull</i>	<i>top-grasp</i>	<i>side-grasp</i>
<i>moved right</i>	0.93	0.00	0.00	0.01	0.01	0.02
<i>moved left</i>	0.00	0.96	0.00	0.02	0.03	0.15
<i>moved forward</i>	0.00	0.00	0.89	0.01	0.01	0.04
<i>pulled</i>	0.00	0.00	0.00	0.87	0.01	0.02
<i>disappeared</i>	0.00	0.00	0.00	0.09	0.00	0.03
<i>grasped</i>	0.00	0.00	0.00	0.00	0.23	0.17
<i>knocked</i>	0.03	0.02	0.08	0.00	0.07	0.10
<i>no effect</i>	0.04	0.02	0.03	0.00	0.64	0.47

gorithm (Kononenko, 1994) to compute the relevancy of each feature to the prediction of the effect labels $f \in \mathcal{F}$ in the data set. Then, using only those features whose (relevancy) weights are larger than zero, we train an SVM for each behavior b_i to learn a mapping of the form $\mathcal{V}_E \rightarrow \mathcal{F}$ as can be seen in Figure 6. In our experiments, the SVM’s, using Radial Basis Functions as kernel functions, were trained using 5-fold cross-validation where, in each iteration, 80% of the data was used for training and the remaining for testing to achieve accuracy values above 90%.

After training, the SVM classifiers denoted by \mathcal{M}_{b_k} can then be used to predict the effect label $f_{o_l}^{b_k}$ of a behavior b_k on a novel object o_l , as sketched in Figure 6(b). This allows the robot to predict the effects along with confidences, that are likely to be generated as a result of applying that behavior.

Table 5 shows the matrix of effect confidences for a single object. For instance, the first column of the matrix shows that, when the robot applies the *push right* behavior on the object, we have 93% confidence that the object will be *pushed-right*, and 3% confidence that the object will be *knocked* and 4% confidence that the object will have no change. All the confidences in this matrix are used as the *affordance vector*.

6 Experimental results

We proposed and evaluated three different models for learning adjectives and nouns. As shown in the lower row of ellipses in Figure 7, these models are essentially SVM classifiers that map their input into adjectives or nouns from one of the following inputs:

- *Entity*: \mathcal{V}_E which consists of the perceptual view of the object obtained from the point cloud data as described in Section 4.3. In this sense, \mathcal{V}_E includes features obtained from the visual appearance of the object and is also referred to as *appearance features* in the rest of the article.

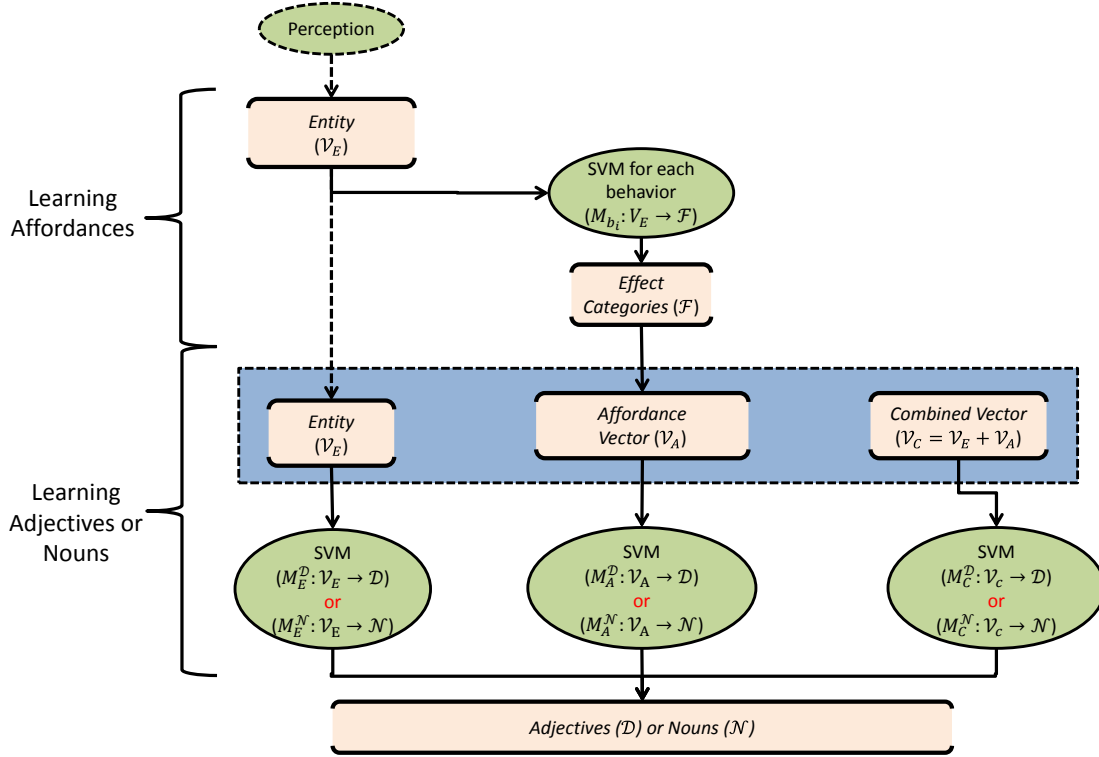


Figure 7: Overview of the system. The robot first learns the affordances of objects and then adjectives and nouns. For learning adjectives and nouns we have four different methods, M_E^* , M_A^* and M_C^* (* denotes either \mathcal{D} or \mathcal{N}), that map to the set of adjectives \mathcal{D} or the set of nouns \mathcal{N} . M_E^* is appearance-based, whereas M_A^* is functional since they are based on objects' affordances. M_C^* , on the other hand, includes both appearance and functional aspects.

- *Affordance vector*: $\mathcal{V}_A = (\hat{f}_1^{b_1}, \dots, \hat{f}_8^{b_1}, \dots, \hat{f}_1^{b_6}, \dots, \hat{f}_8^{b_6})$, where $\hat{f}_i^{b_j}$ is the confidence of behavior b_j producing effect f_i on the object o , as shown in Figure 5. In this sense, \mathcal{V}_A includes features about the affordances of the object and is also referred to as *affordance features* in the rest of the article.
- *Combined vector*: \mathcal{V}_C which is formed by the concatenation of \mathcal{V}_A and \mathcal{V}_E .

6.1 Learning Adjectives

We constructed and evaluated three different models, denoted as M_A^D , M_E^D and M_C^D , based on their input. Each model consisted of three different SVM classifiers for each pair of adjectives.

Figure 8(a) depicts the average, maximum, and minimum predic-

tion accuracies for the different models obtained during the 5-fold cross-validation training. It can be seen that M_A^D performs slightly better than M_E^D and much better than M_C^D . This result indicates that adjectives can be learned better using affordance features rather than using appearance features. This is interesting since affordances are also computed from the appearance of the object. The M_C^D model which used concatenated affordance features with appearance features performed the worst, indicating that inclusion of appearance features provided conflicting information that deteriorated the quality of the generalization for predicting adjectives.

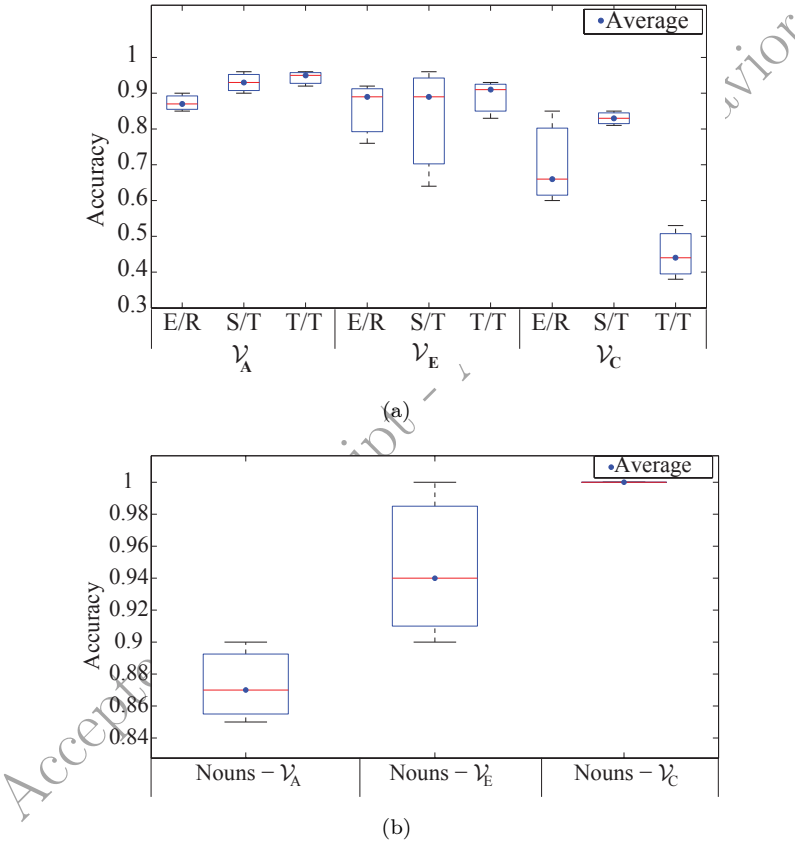


Figure 8: Whisker plot depicting the distribution of prediction accuracies of each model for learning (a) adjectives and (b) nouns. In (a), for the sake of space, we abbreviate adjective groups Edgy/Round, Short/Tall and Thin/Thick with E/R, S/T and T/T, respectively.

6.2 Learning Nouns

We used one SVM classifier for nouns in each model to learn the mapping between $\mathcal{V}_A, \mathcal{V}_E$ and \mathcal{V}_C and nouns. These models are denoted as M_A^N, M_E^N and M_C^N . Figure 8(b) depicts the average, maximum and minimum prediction accuracies for different models obtained during the 5-fold cross-validation training. It can be seen that M_E^N performs better than M_A^N . The result indicates that, unlike the case for adjectives, appearance features are better than affordance features in predicting nouns. Moreover, the concatenation of appearance features with affordance features improves the prediction accuracy for nouns as can be seen for M_C^N .

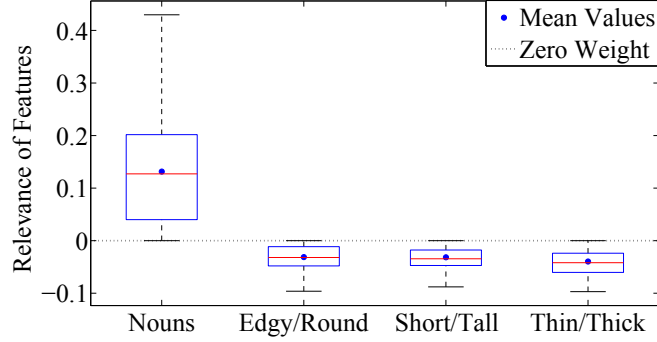
6.3 Adjectives versus Nouns

Figures 8(a) and 8(b), with the significance analysis in Table 6, show that adjectives are better learned with affordance features whereas nouns are better learned from appearance features, indicating an underlying distinction. In order to further investigate this finding, we analyzed the relevancy weights of affordance and appearance features in \mathcal{V}_C for learning adjectives and nouns using the ReliefF algorithm (Kononenko, 1994). In ReliefF, the larger the weight of a feature, the more relevant the feature is for classification. A weight of zero means that the given feature has no contribution in classification with respect to the given labels. On the other hand, a negative weight indicates that the given feature is more likely to cause misclassifications.

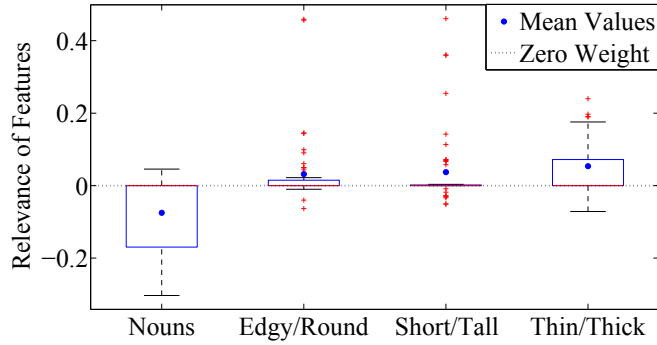
Table 6: χ^2 significance test results for the noun and adjective learning models based on the three representations ($\mathcal{V}_A, \mathcal{V}_E$ and \mathcal{V}_C). The test outcomes with statistical significance ($p \leq 0.05$) are marked in bold. The noun models have 3 degrees of freedom while each adjective pair model has 1 degree of freedom.

	Nouns (df=3)	Edgy/Round (df=1)	Short/Tall (df=1)	Thin/Thick (df=1)
\mathcal{V}_A	0.67	1.21	0.21	1.00
\mathcal{V}_E	0.22	10.28	3.66	2.83
\mathcal{V}_C	0.0	211.76	4.0	2.83

Figure 10 plots the relevancy weights of features (grouped into appearance and affordance) for learning nouns and adjective pairs. It can be seen that perceptual features are more important for nouns than affordance features, but vice versa for adjectives. Figure 9 plots the distribution of relevancy weights for appearance and affordance features for nouns and adjectives proving the distinction between the two types of concepts in language. The plot in (a) shows that appearance features are informative for nouns but detrimental for adjectives. On the other hand, the plot in (b) shows that affordance features are detrimental for nouns but informative for adjectives. The negative relevance weights of appearance features explains the drop in the prediction performance of \mathcal{M}_C^D for learning adjectives.



(a)



(b)

Figure 9: Whisker plot for the distribution of the relevance of features of the appearance vector \mathcal{V}_E (a) and the affordance vector \mathcal{V}_A (b) for nouns and adjectives.

Table 7: Average, maximum and minimum ReliefF weights of each category test.

Categorizations	Appearance Features			Affordance Features		
	Avg.	Max.	Min.	Avg.	Max.	Min.
Nouns	0.13	0.42	0.00	-0.07	0.04	-0.30
Edgy/Round	-0.03	0	-0.09	0.03	0.45	-0.06
Tall/Short	-0.03	0	-0.08	0.03	0.46	-0.05
Thin/Thick	-0.03	0	-0.09	0.05	0.41	-0.07

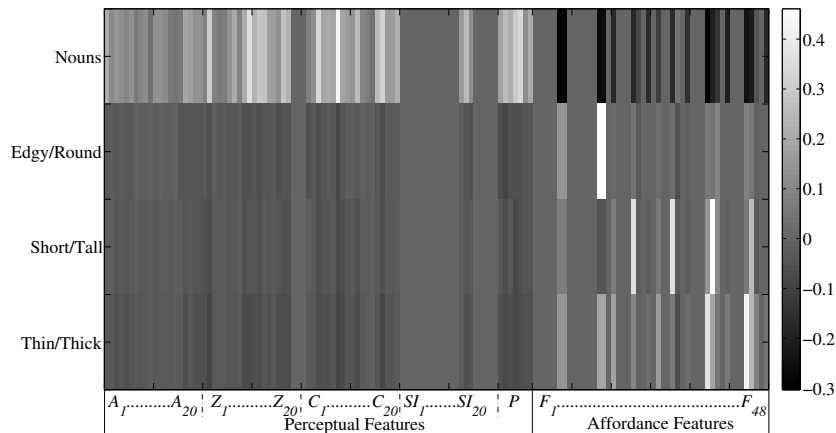


Figure 10: Relevance of the features in the combined vector representation \mathcal{V}_c . The features with higher weights are denoted with a lighter shade of gray. The features are abbreviated as L_i (Azimuth Histograms), Z_i (Zenith Histograms), C_i (Curvature Histograms), SI_i (Shape Index Histograms), P (Position, Orientation, Size and Presence), and A_i (Affordance features).

This significant distinction has important implications regarding the *functional* distinction between nouns and adjectives. Differences between adjectives and nouns have already been pointed out or implied by neuroscientists, psychologists and linguists. For example, recent fMRI recordings regarding which parts of the brain are activated for verbs, nouns and adjectives, show that adjectives are related to actions more than nouns (Liang et al., 2006). This is in line with our findings, since adjectives are better learned with objects’ affordances.

Developmental psychologists have reported that young children have more difficulty learning noun modifying adjectives than nouns (Sandhofer and Smith, 2007). Considering that the adjectives used in our experiments are noun modifying adjectives, one can consider using affordances for learning adjectives as providing a middle-layer of abstraction over the raw sensorispace. In other words, this mid-level abstraction layer increases the performance of the learning of adjectives.

Linguists like Sassoon (2011) hypothesize that adjectives describe object categories along single or few dimensions of feature space whereas nouns use almost all the dimensions. The adjectives used in our experiments conform to this hypothesis since they are dependent only on a few dimensions such as “width”, “height” etc. SVMs thus have more difficulty learning these adjectives directly from appearance because the irrelevant changes in other dimensions distract the learner. However, when affordances are used, they provide a layer of abstraction that captures only information in the relevant dimensions, making the learning of adjectives easier.

6.4 Adjectives and Nouns for Novel Objects

We evaluated the adjective and noun prediction performances of the different models on novel objects in our lab as well as on novel objects from the KIT 3D object database (Kasper et al., 2012).

Figure 11 shows the adjectives and nouns predicted for novel objects by the different models. For adjectives, M_A^D , which uses affordance features as input, makes a single misprediction (describing O_3 as *short*) out of 42 predictions. The other models, M_E^D and M_C^D , that use appearance features as input made 10 and 4 mispredictions in agreement with our earlier results that adjectives are best described by affordances.

For nouns, all the models agree on 12 of the 14 objects, all being correct but one. All models mispredicted O_4 (a plastic hamburger toy) as a *box*, possibly due to its sharp edges. The models disagreed on naming objects O_4 (a plastic bone toy for dogs) and O_6 (a glue bottle). M_E^N and M_C^N named the plastic bone as a cylinder correctly, whereas M_A^N mispredicted it as a *box*. There are no correct noun descriptors for the glue bottle, O_6 , and all predictions are considered as wrong. M_E^N and M_C^N prediction confidences for box prediction was below 50%. M_A^N 's prediction of the object as a *cup* with a confidence of 89% can be considered as bad.

The number of correct noun predictions for all models are close to each other, and a closer look at the confidences is needed. M_E^N 's confidence in the correct predictions is above 86%, whereas its confidence drops to 46% for the glue bottle case. In contrast, M_A^N 's confidence for the correct predictions ranged between 56% and 89%. However it is disappointing to see that the model has high confidence values (94% for the plastic bone and 89% for the glue bottle) for mispredictions.

The performance of M_C^N , which used both appearance and affordance features lies in between the other models. Its confidence values for correct noun predictions are even lower (between 52% and 80%) than the ones obtained from M_A^N . However, it correctly predicts the plastic bone as a cylinder (with 52% confidence) and has low confidence in its prediction for the glue bottle. It confirms our prior results that the inclusion of affordance features taints the model and deteriorates its performance.

6.5 Odd One Out

Finding an odd object in a set is a challenging cognitive benchmarking task in which the subject is presented with a set of objects and asked to mark the item that is least similar to the others. Odd-one out is a feasible way to comment on the robustness of object representation, especially in terms of dealing with unknown objects. However, only few robotics and computer vision researchers have studied this interesting problem.

We used the M_A^D , which has the best adjective prediction performance, for finding an odd object in a set of objects. For this, the robot predicts the adjectives ($a_1^i, a_2^i, a_3^i \in \mathcal{D}$) of each object o_i in a scene with the corresponding prediction confidences p_1^i, p_2^i, p_3^i . The robot then compares the predicted adjectives of an object against others to find the odd one

as:

$$D(o_i) = \sum_{o_j \in \mathcal{S}} \sum_{k=1}^3 d(o_i, o_j, k), \quad (5)$$

where \mathcal{S} is the set of objects in the scene, and $d(\cdot)$ is defined as:

$$d(o_i, o_j, k) = \begin{cases} |p_k^i - p_k^j|, & \text{if } a_k^i = a_k^j \\ p_k^i + p_k^j, & \text{otherwise} \end{cases}. \quad (6)$$

marking the object with the highest $D()$ as odd.

Five sample executions for the odd-one-out task are shown in Figure 12. Note that although there are many possible ways to select an odd object, the robot does so by comparing each object via the predicted adjectives. As can be seen from Figure 11, the robot can choose the odd object in a set by looking at objects’ adjectives which are estimated from the objects’ affordances.

There are a number of related studies on the Odd-One-Out problem. For example, Choi et al. (2010) used probabilistic reasoning over a large database of object categories to form hierarchical context structures. With their methodology, they can extract the odd object from a scene. In their study, an object is determined as “odd” if its pose, scale, scene or occurrence is inconsistent.

Lemaignan et al. (2010) use a common-sense ontology designed for robots to draw useful conceptual information about objects. In their experiments, the robot picks each object and asks a human about the object’s properties until it can successfully match it with available concepts. After all the objects have been identified in the available concepts, the odd object can be selected. Although this study is based on common sense knowledge, the conceptualizations are not linked with the robot’s own sensorimotor experiences.

The study by Sinapov and Stoytchev (2010) has strong parallels with our approach. With a behavior repertoire consisting of *lift*, *shake*, *drop*, *crush* and *push*, they have their robot interact with a large set of objects. Using the sensory feedback data associated with the interaction experiences of each object, they form a matrix to formulate the similarities between the objects. When a group of objects is presented, their robot utilizes this matrix to distinguish the odd object.

7 Conclusion

In this article we studied the links between nouns and adjectives in language with the sensorimotor interactions of a robot. Specifically, an iCub humanoid robot interacted with a set of objects using a repertoire of manipulation behaviors. During these interactions, a human observed the object being manipulated as well as the effect induced. The objects are labeled by three shape-related adjectives and a noun, whereas the effect is labeled as an affordance. The data collected through these interactions is used (i) to learn the affordances of objects, and (ii) to learn a mapping from the affordance and appearance related features of objects to adjectives and nouns.

We proposed and evaluated three different models for the learning of adjectives and nouns through cross-validated training as well as through testing on novel objects. Moreover, we have analyzed the relevancy of affordance and appearance related features for the learning of adjectives and nouns. We have also shown that adjective predictions can be used to solve the odd-one-out task in a successful way on a number of examples.

The results showed that adjectives are better learned using objects' affordances, whereas nouns are better learned using objects' appearance. This distinction is discussed in relation to findings from Psychology and Neuroscience, and Linguistics. Furthermore, our computational results (i) provide evidence, especially relevant to Developmental Psychologists, that learning adjectives and nouns show underlying differences, (ii) support the hypothesis (Sassoon, 2011) that adjectives span fewer feature dimensions than nouns do.

8 Discussion

- We argue that our main contribution lies in the proposed computational framework, as well as analysis method, for studying the links between word categories in language and the sensorimotor interactions of an organism in the environment. The proposed framework allows us to study the conflicts between appearance and affordance (of function) based views of categories in language.
- Obviously the learning of adjectives and nouns is studied in a rather limited context that is mostly determined by the perception and manipulation capabilities of the robot platform. The limited set of manipulation behaviors, although on a par with related studies, forced us to use only the basic physical affordances of objects. The use of point cloud data (rather than visual data) allowed us to use features that are more related to the physical affordances of an object. However, it also limited our study to use shape-related adjectives and nouns. Moreover, the number of physical interactions that can be obtained from a physical robot is rather limited due to the physical unreliability and the cost of these platforms.
- The limited context of our study prohibits us from making general claims about the learning of adjectives and nouns in language. Nevertheless, the argument that shape-related adjectives are mostly based on the affordances of objects is likely to hold. However, this claim does not apply to all adjectives. An obvious example is color adjectives which are based on the appearance of objects and have nothing to do with their affordances.
- We would like to point out that affordances depend on the behavioral interactions of the robot. Although the proposed method is independent of the robot platform, the use of a human-like robot and human-like behaviors is likely to develop concepts similar on robots to the ones that humans have.
- The use of human labeling is a must for creating human-like categories or concepts in the robot. Self-organized categorization of

interaction data is highly dependent on the perceptual representation of the robot, and can generate categorizations that may not necessarily coincide with the linguistic concepts that humans have.

- Finally, we would like to point to a complimentary study (Kalkan et al., 2010) where we studied the linking of verbs to the sensorimotor interactions of the robot through affordances.

Acknowledgments

This work is partially funded by the EU projects ROSSI (FP7-ICT-216125), and RobotCub (FP6-ICT-004370), and by TÜBİTAK (Turkish Scientific and Technical Council) through projects no 109E033 and 111E287.

References

- Aksoy, E. E., Abramov, A., Worgotter, F., and Dellen, B. (2010). Categorizing object-action relations from semantic scene graphs. *IEEE International Conference on Robotics and Automation*, pages 398 – 405.
- Atıl, I., Dağ, N., Kalkan, S., and Şahin, E. (2010). Affordances and emergence of concepts. *10th International Conference on Epigenetic Robotics*.
- Baker, M. C. (2003). *Lexical categories: Verbs, nouns and adjectives*, volume 102. Cambridge University Press.
- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.
- Bergquist, T., Schenck, C., Ohiri, U., Sinapov, J., Griffith, S., and Stoytchev, A. (2009). Interactive object recognition using proprioceptive feedback. *IROS Workshop: Semantic Perception for Mobile Manipulation*.
- Borghini, A. M. (2007). Object concepts and embodiment: Why sensorimotor and cognitive processes cannot be separated. *La nuova critica*, 15(4):447–472.
- Borghini, A. M., Di Ferdinando, A., and Parisi, D. (2002). The role of perception and action in object categorization. In J.A. Bullinaria & W. Lowe (Eds), *Connectionist Models of Cognition and Perception*. Singapore: World Scientific, pages 40–50.
- Borghini, A. M., Flumini, A., Cimatti, F., Marocco, D., and Scorolli, C. (2011). Manipulating objects and telling words: a study on concrete and abstract words acquisition. *Frontiers in psychology*, 2.
- Bub, D., Masson, M., and Cree, G. (2008). Evocation of functional and volumetric gestural knowledge by objects and words. *Cognition*, 106(1):27–58.

- Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2):139–151.
- Cangelosi, A. and Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4(1):117–142.
- Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., Fadiga, L., Wrede, B., Rohlfing, K., Tuci, E., Dautenhahn, K., Saunders, J., and Zeschel, A. (2010). Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195.
- Cangelosi, A. and Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, 89(2):401–408.
- Cangelosi, A. and Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive science*, 30(4):673–689.
- Carbonetto, P. and de Freitas, N. (2003). Why can't jose read? the problem of learning semantic associations in a robot environment. In *The HLT-NAACL 2003 Workshop on Learning word meaning from non-linguistic data*, pages 54–61.
- Chauhan, A. and Lopes, L. S. (2011). Using spoken words to guide open-ended category formation. *Cognitive Processing*, 12(4):341–354.
- Chella, A., Dindo, H., and Zambuto, D. (2009). Grounded human-robot interaction. In *Biologically Inspired Cognitive Architectures: 2009 AAAI Fall Symposium Series*.
- Choi, M. J., Lim, J. J., Torralba, A., and Willsky, A. S. (2010). Exploiting hierarchical context on a large database of object categories. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 129 – 136.
- Crutch, S. J. and Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.
- Şahin, E., Çakmak, M., Doğar, M. R., Uğur, E., and Üçoluk, G. (2007). To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior*, 15(4):447–472.
- Dağ, N., Atıl, I., Kalkan, S., and Şahin, E. (2010). Learning affordances for categorizing objects and their properties. *IEEE International Conference on Pattern Recognition (ICPR)*.

- Dindo, H. and Zambuto, D. (2010). A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 790–796. IEEE.
- Fischer, M. and Zwaan, R. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*, 61(6):825–850.
- Gallese, V. and Lakoff, G. (2005). The brain’s concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, 22(3):455–479.
- Gibson, E. J. (2000). Perceptual learning in development: Some basic concepts. *Ecological Psychology*, 12(4):295–302.
- Gibson, J. J. (1986). *The Ecological Approach to visual perception*. Lawrence Erlbaum Associates.
- Glenberg, A. and Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558.
- Glenberg, A. and Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43(3):379–401.
- Glenberg, A., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., and Buccino, G. (2008). Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology*, 61(6):905–919.
- Glenberg, A. M. and Gallese, V. (2011). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48(7):905–922.
- Gold, K., Donic, M., Crick, C., and Scassellati, B. (2009). Robotic vocabulary building using extension inference and implicit contrast. *Artificial Intelligence*, 173(1):145–166.
- Goodale, M. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25.
- Haazebroek, P., van Dantzig, S., and Hommel, B. (2011). A computational model of perception and action for cognitive robotics. *Cognitive Processing*, 12(4):355–365.
- Harnad, S. (1990). The symbol grounding problem. *Physica*, D(42):335–346.
- Helbig, H. B., Graf, M., and Kiefer, M. (2006). The role of action representations in visual object recognition. *Experimental Brain Research*, 174(2):221–228.

- Jax, S. and Buxbaum, L. (2010). Response interference between functional and structural actions linked to the same familiar object. *Cognition*, 115(2):350–355.
- Kalkan, S., Dag, N., Yuruten, O., Borghi, A. M., and Sahin, E. (in press). Verb concepts from affordances. *Interaction Studies Journal*.
- Kasper, A., Xue, Z., and Dillmann, R. (2012). The kit object models web database: An object model database for object recognition, localization and manipulation in service robotics. *The International Journal of Robotics Research*, 31:927–934.
- Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. *European Conference on Machine Learning*, 784:171–182.
- Krüger, N., Geib, C., Piater, J., Petrick, R., Steedman, M., Worgotter, F., Ude, A., Asfour, T., Kraft, D., Omrcen, D., Agostini, A., and Dillmann, R. (2011). Object–action complexes: Grounded abstractions of sensory–motor processes. *Robotics and Autonomous Systems*, 59(10):740–757.
- Lemaignan, S., Ros, R., Mösenlechner, L., Alami, R., and Beetz, M. (2010). Oro, a knowledge management platform for cognitive architectures in robotics. *International Conference on Intelligent Robots and Systems*, pages 3548 – 3553.
- Liang, D., Yang, Y., Feng, S., and Li, J. (2006). A fmri study on modifiers of noun by nouns, verbs and adjectives in chinese. *Applied Linguistics*, 4:13.
- McMahon, I., Chu, V., Rifano, L., McDonald, C. G., He, Q., Perez-Tejada, J. M., Arrigo, M., Fitter, N., Nappo, J. C., and Darrell, T. (2012). Robotic learning of haptic adjectives through physical interaction. *IROS workshop on Advances in Tactile Sensing and Touch based Human-Robot Interaction*.
- Metta, G. and Fitzpatrick, P. (2003). Better vision through manipulation. *Adaptive Behavior*, 11(2):109–128.
- Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pages 50–56.
- Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008a). Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26.
- Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008b). Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26.

- Montesano, L., Lopes, M., Melo, F., Bernardino, A., and Santos-Victor, J. (2009). A computational model of object affordances. In *Advances In Cognitive Systems*. IET.
- Morse, A. F., Baxter, P., Belpaeme, T., Smith, L. B., and Cangelosi, A. (2011). The power of words. *Int. Conference on Epigenetic Robotics*.
- Pellicano, A., Iani, C., Borghi, A. M., Rubichi, S., and Nicoletti, R. (2010). Simon-like and functional affordance effects with tools: The effects of object perceptual discrimination and object action state. *The Quarterly Journal of Experimental Psychology*, 63(11):2190–2201.
- Petrosino, A. and Gold, K. (2010). Toward fast mapping for robot adjective learning. In *Dialog with Robots: 2010 AAAI Fall Symposium Series*.
- Pexman, P. M., Hargreaves, I. S., Edwards, J. D., Henry, L. C., and Goodyear, B. G. (2007). Neural correlates of concreteness in semantic categorization. *Journal of Cognitive Neuroscience*, 19(8):1407–1419.
- Rusu, R. B. and Cousins, S. (2011). 3d is here: Point cloud library (pcl). *Library*, 26(2):1–4.
- Salvi, G., Montesano, L., Bernardino, A., and Santos-Victor, J. (2012). Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems Man, and Cybernetics Part B Cybernetics*, 42(3):660–671.
- Sandhofer, C. and Smith, L. B. (2007). Learning adjectives in the real world: How learning nouns impedes learning adjectives. *Language Learning and Development*, 3(3):233–267.
- Sassoon, G. (2011). Adjectival vs. nominal categorization processes: The rule vs. similarity hypothesis. *Belgian Journal of Linguistics*, 25(1):104–147.
- Sinapov, J. and Stoytchev, A. (2010). The odd one out task: Toward an intelligence test for robots. *9th IEEE International Conference on Development and Learning (ICDL)*, pages 126–131.
- Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, 7(7):308–312.
- Sugita, Y., Tani, J., and Butz, M. V. (2011). Simultaneously emerging braintenberg codes and compositionality. *Adaptive Behavior*, 19(5):295–316.
- Sun, J., Moore, J., Bobick, A., and Rehg, J. (2010). Learning Visual Object Categories for Robot Affordance Prediction. *The International Journal of Robotics Research*, 29(2-3):174–197.
- Wu, C. and Aghajan, H. (2009). Using context with statistical relational models: object recognition from observing user activity in home environment. *Workshop on Use of Context in Vision*.

- Yu, C. and Ballard, D. H. (2004). On the integration of grounding language and learning objects. *19th Int. Conf. on Artificial Intelligence*, pages 488–493.
- Yuruten, O., Uyanik, K. F., Caliskan, Y., Bozcuoglu A. K., Sahin, E., and Kalkan, S. (2012). Development of adjective and noun concepts from affordances on the icub humanoid robot. *12th International Conference on Adaptive Behaviour (SAB)*.
- Zwaan, R. and Taylor, L. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology-General*, 135(1):1–11.

Accepted Manuscript - Adaptive Behavior















Object	Adjectives			Nouns		
	M_A^D	M_E^D	M_C^D	M_A^N	M_E^N	M_C^N
O_1 	edgy (54%) short (97%) thin (59%)	edgy (89%) short (55%) thin (52%)	edgy (60%) short (80%) thin (52%)	box (74%)	box (97%)	box (56%)
O_2 	round (77%) short (77%) thin (89%)	edgy (79%) short (58%) thin (67%)	round (65%) short (68%) thin (62%)	ball (83%)	ball (97%)	ball (80%)
O_3 	edgy (63%) short (94%) thin (96%)	edgy (64%) tall (67%) thin (84%)	edgy (60%) tall (68%) thin (80%)	cyl. (87%)	cyl. (95%)	cyl. (60%)
O_4 	round (84%) short (98%) thick (91%)	round (77%) short (68%) thin (62%)	round (75%) short (71%) thick (51%)	box (94%)	cyl. (86%)	cyl. (52%)
O_5 	round (84%) short (97%) thick (95%)	round (89%) short (67%) thick (58%)	round (80%) short (66%) thick (54%)	box (89%)	box (94%)	box (62%)
O_6 	edgy (84%) short (98%) thin (92%)	edgy (79%) tall (55%) thick (62%)	edgy (75%) short (65%) thick (52%)	cup (89%)	box (46%)	box (45%)
O_7 	edgy (62%) short (98%) thick (78%)	round (84%) short (54%) thick (68%)	edgy (60%) short (56%) thick (66%)	box (89%)	box (93%)	box (64%)
O_8 	round (72%) short (98%) thick (79%)	edgy (89%) short (67%) thick (52%)	round (62%) short (69%) thick (53%)	cup (89%)	cup (98%)	cup (61%)
K_1 	round (60%) tall (97%) thick (76%)	edgy (92%) tall (100%) thick (96%)	round (60%) tall (80%) thin (52%)	cyl. (61%)	cyl. (98%)	cyl. (56%)
K_2 	round (55%) tall (96%) thick (72%)	edgy (90%) tall (98%) thick (91%)	round (62%) tall (82%) thick (54%)	cyl. (56%)	cyl. (98%)	cyl. (58%)
K_3 	edgy (55%) tall (97%) thin (72%)	edgy (92%) tall (95%) thin (93%)	edgy (92%) tall (79%) thin (81%)	box (58%)	box (97%)	box (59%)
K_4 	round (58%) tall (98%) thick (87%)	edgy (76%) tall (100%) thick (86%)	edgy (82%) tall (83%) thick (70%)	cup (61%)	cup (96%)	cup (68%)
K_5 	round (55%) tall (95%) thick (71%)	edgy (76%) tall (98%) thick (94%)	edgy (80%) tall (80%) thick (52%)	cup (56%)	cup (98%)	cup (56%)
K_6 	edgy (59%) tall (92%) thick (92%)	edgy (83%) tall (96%) thick (90%)	edgy (62%) tall (78%) thick (52%)	box (56%)	box (99%)	box (62%)

Figure 11: Predicted adjectives and nouns, along with confidence values in parenthesis, for novel objects. Objects labeled as O_i are picked within our lab, where as objects labeled as K_i are picked from the KIT 3D object database (Kasper et al., 2012). Bold labels denote correct classifications.











Scene	Selected Odd Object	Predicted Adjectives			
	 (Paint Box)	Object Properties			
		Spray edgy (63%) short (94%) thin (96%)	Glue edgy (84%) short (98%) thin (92%)	Paint Box edgy (62%) short (98%) thick (78%)	Lamp Box edgy (54%) short (97%) thin (59%)
	 (Ball)	Plastic Cup round (72%) short (98%) thick (79%)	Hamburger round (84%) short (97%) thick (95%)	Ball round (77%) short (91%) thin (67%)	Toy Bone round (84%) short (98%) thick (91%)
	 (Ball)	Lamp Box edgy (54%) short (97%) thin (59%)	Glue edgy (84%) short (98%) thin (92%)	Ball round (77%) short (91%) thin (67%)	Spray edgy (63%) short (94%) thin (96%)
	 (Wooden Cyl.)	Ball round (77%) short (91%) thin (67%)	Wooden Cyl. round (80%) tall (82%) thin (62%)	Small Cyl. round (86%) short (92%) thin (90%)	Pencil Box round (84%) short (96%) thin (72%)
	 (Wooden Cyl.)	Paint Box edgy (84%) short (98%) thin (92%)	Hamburger round (84%) short (97%) thick (95%)	Wooden Cyl. round (80%) tall (82%) thin (62%)	Ball round (77%) short (91%) thin (67%)

Figure 12: Odd-One-Out experiments with the M_A^D model. The robot successfully detects the adjectives for each object, then chooses the odd object. The selected odd objects and the adjectives that are most effective for determining the oddness are marked in bold.