Towards Understanding Object-Directed Actions: A Generative Model for Grounding Syntactic Categories of Speech through Visual Perception

Amir Aly, Tadahiro Taniguchi

Towards Understanding Object-Directed Actions: A Generative Model for Grounding Syntactic Categories of Speech through Visual Perception

Amir Aly[1] and Tadahiro Taniguchi[2]

***Abstract*— Creating successful human-robot collaboration requires robots to have high-level cognitive functions that could allow them to understand human language and actions in space. To meet this target, an elusive challenge that we address in this paper is to understand object-directed actions through grounding language based on visual cues representing the dynamics of human actions on objects, object characteristics (color and geometry), and spatial relationships between objects in a tabletop scene. The proposed probabilistic framework investigates unsupervised Part-of-Speech (POS) tagging to determine syntactic categories of words so as to infer grammatical structure of language. The dynamics of object-directed actions are characterized through the locations of the human arm joints – modeled on a Hidden Markov Model (HMM) – while manipulating objects, in addition to those of objects represented in 3D point clouds. These corresponding point clouds to segmented objects encode geometric features and spatial semantics of referents and landmarks in the environment. The proposed Bayesian learning model is successfully evaluated through interaction experiments between a human user and Toyota HSR robot in space.**

I. INTRODUCTION

Robots are moving to the center of the human social environment; they are autonomously navigating into spaces and collaborating with human users in different tasks. This requires robots to understand object-directed actions through verbal instructions from human users, which embraces different challenges, like learning the associated dynamics to action verbs, object features, and spatial relationships between referents and landmarks in the environment [13, 24].

Grounding language through perception is a key challenge both in artificial intelligence and cognitive science. The "Symbol Grounding" problem was first defined by Harnad [18], which indicates assigning a meaning to every meaningless symbol (e.g., a new word) in a structure through interaction with the surroundings. Tanenhaus et al. [43] investigated the important effect of visual cues on understanding spoken language. Roy [32] proposed a model for visually grounding words that describe a scene through a "show-and-tell" training procedure. Roy et al. [33] discussed an architecture that can provide perceptual and affordance representations of words so as to make robots able to better engage in interaction. Salvi et al. [35] developed a probabilistic affordance model to learn word meanings in terms of actions and object features. Matuszek et al. [26] proposed a probabilistic joint-learning framework that employs categorial grammar to develop compositional meaning representations of language and physical objects in the environment.

Grounding action verbs (e.g., push, raise, and touch) in sensorimotor behavior has been fueled by the ambition of making robots able to learn the associated motor behavior to the verbs [13]. Siskind [37] developed a computational model to ground action verbs (e.g., pick-up, put-down, and move) through analyzing visual scenes in short image sequences for a human performing actions. Tani et al. [44] used a Recurrent Neural Network with Parametric Biases (RNNPB)-based model to learn and generate different types of dynamic behavioral patterns associated with action verbs. Cohen et al. [9] proposed a semantic representation of verbs to ground action verbs (e.g., push and hit) based on dynamic features representing distance, velocity, and transfer of energy between two physical entities in interaction. Marocco et al. [25] proposed a framework for grounding action words through sensorimotor interaction with the environment.

Grounding spatial concepts has attracted wide attention of researchers in artificial intelligence and cognitive science, and has been extensively studied during the last years. Regier [31] discussed a connectionist learning model that could understand spatial concepts and events of objects in a movie. Cangelosi et al. [6] developed a computational connectionist model that employs an Elman network to encode - in a neural representation - the dynamics of a visual scene. However, these previous approaches do not investigate spatial concepts grounding through a language-based analysis. On the other hand, Tellex et al. [47] proposed a probabilistic learning framework for grounding spatial relationships in natural language instructions within an open space. Dawson et al. [11] used a generative probabilistic model to understand utterances that represent spatial relationships between referents and landmarks in a virtual environment. Guadarrama et al. [17] proposed an interesting probabilistic framework for grounding spatial relationships between objects in space through visual perception. This system includes different modules with different functionalities, such as: object segmentation, action-template matching that associates an utterance to manually constructed templates representing specific actions, and supervised syntactic parsing that segments parts of speech (e.g., nouns of objects) from input texts.

Despite these interesting studies, among many others in the related literature, that discuss language grounding through visual perception, they have not investigated tagging parts of speech through an unsupervised approach so as to infer grammatical structure of speech and the meaning of a

[1]Amir Aly is a senior researcher at the Emergent Systems Laboratory, Ritsumeikan University, Japan amir.aly@em.ci.ritsumei.ac.jp

[2]Tadahiro Taniguchi is a full professor at the Emergent Systems Laboratory, Ritsumeikan University, Japan taniguchi@em.ci.ritsumei.ac.jp

sentence in a developmentally plausible manner. This could help in shedding light on unsupervised grammar induction with a view to making a robot able to understand grammatical dependencies between words in an instruction, which is one of the future topics to address following, and based on, the current study [19]. Furthermore, this approach allows for studying the challenging "Symbol Emergence" problem (i.e., the mechanism of symbols bottom-up development) [45], which attracts wide attention of researchers recently [29], especially after Steels [40] argued that the "Symbol Grounding" problem had been, conceptually, solved.

In this study, we propose an extended computational model of Aly et al. [1] for grounding action verbs, spatial concepts, and object characteristics (color and geometry) of language, with syntactic information, through visual perception, which has not been sufficiently addressed in the literature through a similar global approach that could allow a robot to understand human instructions autonomously. The rest of the paper is organized as follows: Section (II) presents a general description of the system architecture, Sections (III, IV, V) illustrate, in detail, the different subsystems of the framework, Section (VI) illustrates the experimental design, Section (VII) provides a description of the experimental results, and finally, Section (VIII) concludes the paper.

II. SYSTEM ARCHITECTURE

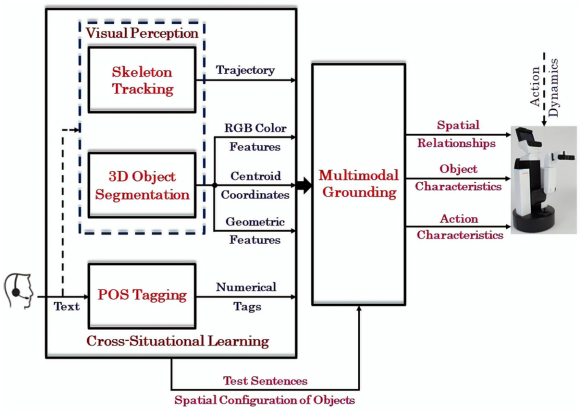


Fig. 1: Overview of the language grounding system

The integrated system in this research study is composed of five different subsystems, as indicated in Figure (1): (1) Speech recognition system (Google HTML5 API speech recognition toolkit) that recognizes human tutor's instructions during interaction with Toyota HSR robot[1], (2) Skeleton tracking and modeling system, which tracks the locations of arm joints while manipulating objects, and models them on a Hidden Markov Model (HMM), (3) 3D Object segmentation system, which segments major planes and objects of the scene into point clouds in order to determine their spatial relationships, in addition to their color and geometric characteristics, (4) Part-of-Speech (POS) tagging system, which marks words of the recognized sentences with syntactic attributes through an unsupervised approach (i.e., *it marks a word sequence with numerical tags without employing any pre-tagged corpus*), and finally (5) Probabilistic learning model for grounding action verbs, spatial concepts, and object characteristics of language through perception.

This model is trained offline though cross-situational learning [38] (Section VI). In case of a test sentence describing a spatial configuration of objects of different colors and shapes, the model determines the spatial relationship between the referent and the landmark so as to allow the robot to perform the required action. The subsystems of this framework are explained, in detail, in the following sections.

III. VISUAL PERCEPTION SYSTEM

A. Skeleton Tracking and Trajectory Modeling

Analyzing and modeling human body motion has been a rich topic for research in the related literature [46]. Ogawara et al. [28] developed a probabilistic framework for learning object manipulation from observation through modeling relative trajectories between objects. Inamura et al. [20] proposed a framework for behavior recognition and generation (e.g., swinging and walking). Sugiura et al. [42] proposed a probabilistic model for learning object manipulation (e.g., place on and jump over) through which the robot learns motion trajectory by demonstration. These approaches, among many others, employ Hidden Markov Models (HMMs) to model action trajectories (i.e., assigning a hidden state sequence to a time-series observation sequence) so as to allow the robot to appropriately learn and generate actions without temporal constraints. In this study, we use a left-to-right HMM model to project the tracked (x, y, z) coordinates of the human arm joints into states while manipulating objects (Figure 2) [20].

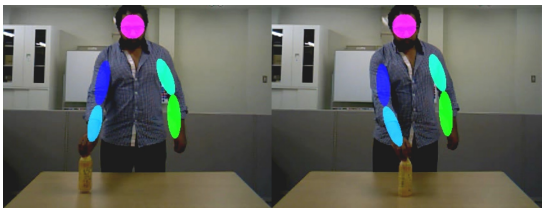


Fig. 2: Human body tracking during object manipulation[2]

The HMM-based gesture model (Figure 3) [30] uses position coordinates of the right-arm joints (i.e., shoulder, elbow, and hand) (transformed into the local coordinate system of the referent) as the observation sequence $\mathrm{P} = (p_1, \ldots, p_m)$ with a hidden state sequence $\mathrm{Q} = (q_1, \ldots, q_n)$. The transition probability between states is $\mathrm{A} = \{a_{ij}\} = \mathbb{P}(q_j|q_i)$, while the emission probability is $\mathrm{B} = \{b_{ij}\} = \mathbb{P}(p_j|q_i)$. Five HMM models have been employed in order to represent the five action verbs of focus in this study (Section VI). During the cross-situational learning phase, each HMM model is trained using the Expectation-Maximization (EM) algorithm [12] on the coordinates of the human arm joints while

[1]It is an off-the-shelf module that will not be highlighted in this study.

[2]The 3D tracking model employs the natural interaction OpenNI2 SDK and the middleware NITE2.

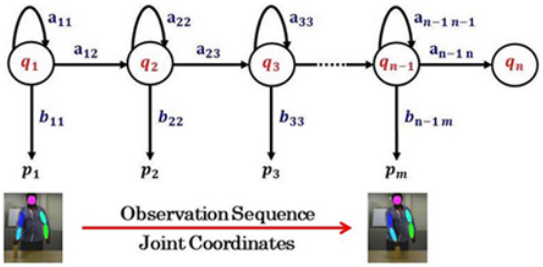


Fig. 3: Left-to-right HMM-based gesture model

doing a specific action in different ways and contexts. The evaluation probabilities of joint coordinates through each HMM model are used to represent actions as observations in the probabilistic learning model (Section V).

B. Unsupervised Object Segmentation into 3D Point Cloud

Different approaches for object segmentation have been investigated in the literature of computer vision. Jiang et al. [21] proposed an edge-based method for calculating the boundaries of point cloud regions through detecting the points with big variance in intensity. Koster and Spann [23] investigated clustering nearby points (i.e., regions) of similar properties using neighboring information. Schnabel et al. [36] discussed a model-based method for clustering points based on geometric criteria, consequently, points with similar mathematical representations create one segment. This last model is employed in this study in order to segment objects lying in a plane into separate point clouds in an unsupervised way [10]. Unlike the previously highlighted approaches, among many others in the literature, this model-based approach is reliable, fast, and does not require much prior knowledge about the environment, such as object models and the number of regions to process.

The employed visual perception model in the framework detects the major plane in the scene (i.e., tabletop or floor) using the RANSAC algorithm [14], and tracks it through successive frames. After calculating the representative equation of the major horizontal plane, orthogonal planes to the major plane, which are in contact with at least one image border, are considered as wall planes. The remaining points in the processed point cloud (out of the corresponding points to the detected horizontal and orthogonal planes) are voxelized and clustered into separate blobs that represent object candidates. Last but not least, the algorithm filters out blobs with very close centroids to orthogonal planes or those at a border of the depth image, in addition to very small and large blobs. Figure (4) illustrates the segmentation output of objects on a tabletop in different spatial layouts.

Both objects and the tabletop plane are segmented into separate point clouds, whose centroids represent their (x, y, z) coordinates with respect to the robot camera. Each point cloud is characterized in terms of its RGB color histogram, in addition to the Viewpoint Feature Histogram (VFH) descriptor (invariant to pose and scale) [34] that efficiently represents the geometry and viewpoint of an object. After calculating the locations of objects on the tabletop, the robot uses a probabilistic learning model (Section V) to ground spatial concepts and object characteristics through cross-situational learning [38] (Section VI).

IV. Unsupervised Part-of-Speech Tagging

Part-of-Speech (POS) tagging is the problem of marking a word sequence with syntactic attributes. The rich literature of computational linguistics has shown different approaches towards grammatically tagging words. Supervised tagging methods use pre-tagged training corpora to set up tagging dictionaries that indicate possible tags of words. These dictionaries are used to tag test words through appropriate stochastic or rule-based models [4, 8]. However, these supervised models would not be capable of tagging new words, which requires using complete dictionaries of language. Semi-supervised tagging methods do not employ pre-tagged large corpora, where they estimate appropriate tags for new word sequences using statistical models [48]. Unsupervised tagging methods do not require any training corpus, where they assign syntactic tags to words using rule-based or statistical models [5, 7]. Therefore, employing this unsupervised approach (i.e., *without using any tagging dictionary*[3]) in our model could help in creating a developmentally plausible system for grounding parts of speech through perception.

A Part-of-Speech (POS) tagging model assigns the syntactic tag $\mathrm{T} = (t_1, \ldots, t_n)$ to the word sequence $\mathrm{w} = (\omega_1, \ldots, \omega_n)$. The first-order Hidden Markov Model (HMM) uses tags as hidden states and words as observations (Figure 5) [15]. The probability distribution of tag states for the word sequence w is defined as follows (where the transition probability to tag t_i is conditioned on tag t_{i-1} to represent the intuitive grammar of language, like having a noun following a determiner):

$$\mathbb{P}(t_1, \ldots, t_n) = \prod_{i=1}^{n} \mathbb{P}(t_i \,|\, t_{i-1}) \tag{1}$$

Emission distributions of tags over words are defined by the probability $\mathbb{P}(\omega_i | t_i)$ of word ω_i conditioned on tag t_i. The transition and emission parameters (ϕ, θ) of the model are characterized by multinomial distributions with Dirichlet priors $(\alpha_\phi, \alpha_\theta)$ (where K denotes the number of tag states):

$$\begin{aligned} t_i \,\big|\, t_{i-1} = t \sim \mathit{Mult}\,(\phi_t) \quad , \quad \phi_t \,\big|\, \alpha_\phi \sim \mathit{Dir}\,(\alpha_\phi) \\ \omega_i \,\big|\, t_i = t \sim \mathit{Mult}\,(\theta_t) \quad , \quad \theta_t \,\big|\, \alpha_\theta \sim \mathit{Dir}\,(\alpha_\theta) \end{aligned} \tag{2}$$

Having an unannotated training corpus with a set of m sentences $\mathrm{W} = \{w_1, \ldots, w_m\}$, the POS tagging model assigns the most likely numerical tag set $\mathrm{T} = \{T_1, \ldots, T_m\}$ for every sentence in the corpus so as to maximize the expression:

$$\begin{aligned} \mathbb{P}(T, W) = \prod_{(T,w) \in (T,W)} \Big(\mathbb{P}(T, w \,|\, \phi, \theta) \Big) = \\ \prod_{(T,w) \in (T,W)} \Big(\prod_{i=1}^{n} \mathbb{P}(t_i \,|\, t_{i-1}, \phi_t) \mathbb{P}(\omega_i \,|\, t_i, \theta_t) \Big) \end{aligned} \tag{3}$$

[3] For example, the POS tagging system could mark every word in the following sentence with the numerical tags: (Push, **1**) (the, **5**) (Red, **2**) (Bottle, **4**) (Near, **6**) (the, **5**) (Cup, **4**), and the learning model will try to ground the meaning of these numerical tags through perception.

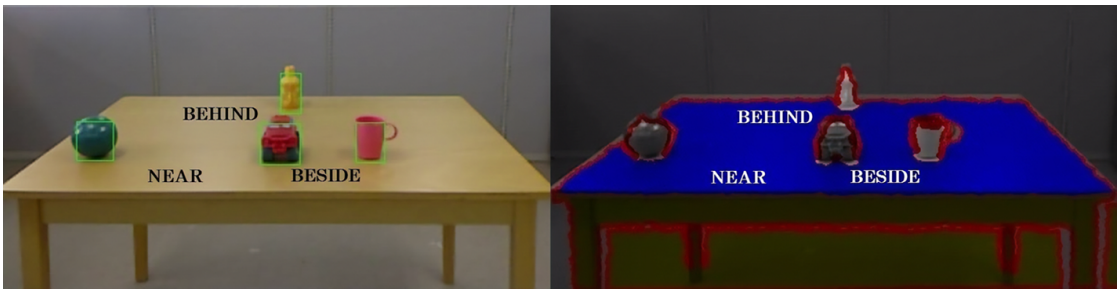


Fig. 4: Different spatial concepts between segmented objects on a tabletop represented through 3D point cloud information

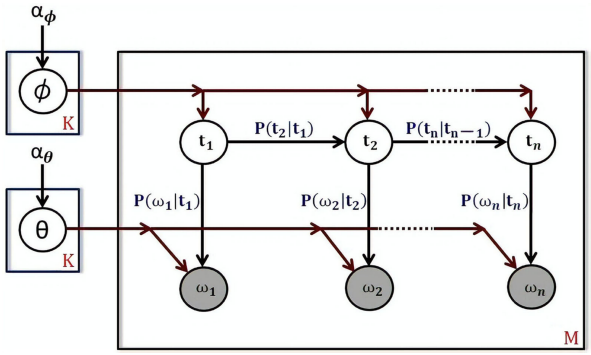


Fig. 5: Graphical representation of the POS tagging model

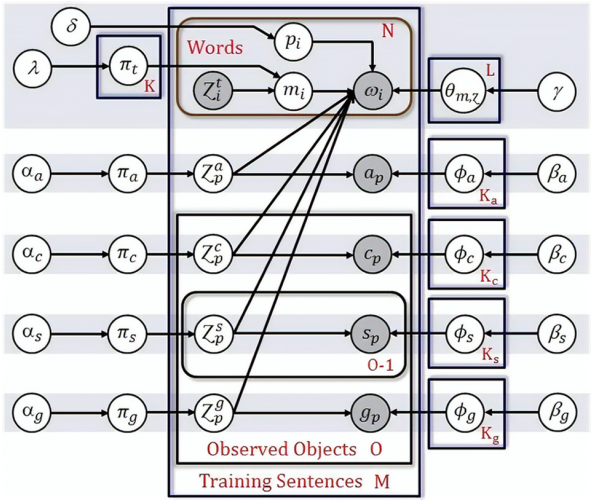


Fig. 6: Graphical representation of the generative model. The index "i" denotes the order of words and their syntactic tags.

The latent tag variables are inferred using the Gibbs sampling algorithm [27], which generates a set of samples from the posterior distribution $\mathbb{P}(T \mid W)$ expressed as follows (where "-i" denotes all samples except the i-th sample):

$$\mathbb{P}(T_i, T^{(i)} \mid T_{-i}, W, T^{(-i)}, w, \alpha_\phi, \alpha_\theta) \tag{4}$$

V. Multimodal Probabilistic Generative Model

The process of grounding action verbs, spatial relationships between referents and landmarks (i.e., between two objects or between an object and the tabletop), and object characteristics (color and geometry) of language through visual perception employs the multimodal Bayesian generative model illustrated in Figure (6). The observed state ω_i represents every word in the sequence $\mathrm{w} = (\omega_1, \ldots, \omega_n)$, and the observed state $\mathcal{Z}_i^t$ represents syntactic tags of words (Section IV). The observed state a_p represents the HMM-modeled locations of arm joints while manipulating objects (Section III-A). The observed state c_p represents the RGB color characteristics (i.e., color histograms) of O observed objects (Section III-B). The observed state s_p represents a spatial layout of a referent A and a landmark B expressed through their centroid coordinates (i.e., the state $s_p = f(P_{A_{x,y,z}}, P_{B_{x,y,z}})$) (Section III-B). The observed state g_p represents the geometric characteristics of O observed objects expressed through the VFH descriptor (Section III-B). For a spatial relationship between a referent and a landmark, the potential dyadic relationships between them could be expressed as follows: *Observed Objects* $O \times (O-1)$ (i.e., Object A $\leftrightarrows$ Object B) (Figure 6). Table (I) summarizes the definitions of the learning model parameters [2]. The probabilistic distributions characterizing the different channels of the model are defined as follows (where *Dir* denotes a Dirichlet distribution, *GIW* denotes a Gaussian Inverse-Wishart distribution, *Cat* denotes a categorical distribution, and *Gauss* denotes a multivariate Gaussian distribution):

TABLE I: Definitions of the generative model parameters in the different modalities.

Parameter	Definition
δ	Hyperparameter of the distribution p_i
p_i	Index of spatial relationship (Object A $\leftrightarrows$ Object B) of each word
λ	Hyperparameter of the distribution π_t
m_i	Index of word modality $\in$ {Action, Color, Layout, Geometry, Others}
γ	Hyperparameter of the distribution $\theta_{m,z}$
L	Number of word distribution categories = $K_a + K_c + K_s + K_g + 1$
$\theta_{m,z}$	Word distribution over modalities
α_a	Hyperparameter of the distribution π_a
β_a	Hyperparameter of the distribution ϕ_a
K_a	Number of categories in the action modality
α_c	Hyperparameter of the distribution π_c
β_c	Hyperparameter of the distribution ϕ_c
K_c	Number of categories in the object color modality
α_s	Hyperparameter of the distribution π_s
β_s	Hyperparameter of the distribution ϕ_s
K_s	Number of categories in the spatial layout modality
α_g	Hyperparameter of the distribution π_g
β_g	Hyperparameter of the distribution ϕ_g
K_g	Number of categories in the object geometry modality
$\mathcal{Z}_p^a$	Index of action categories
$\mathcal{Z}_p^c$	Index of object color categories
$\mathcal{Z}_p^s$	Index of spatial layout categories
$\mathcal{Z}_p^g$	Index of object geometry categories

$$
\begin{cases}
\theta_{m,z_{L_1}} & \sim & Dir\,(\gamma) \quad , & \mathrm{L}_1 = (1,\ldots,\mathrm{L}) \\
\phi_{a_{K_1}} & \sim & GIW\,(\beta_a)\,, & \mathrm{K}_1 = (1,\ldots,\mathrm{K}_a) \\
\phi_{c_{K_2}} & \sim & GIW\,(\beta_c)\,, & \mathrm{K}_2 = (1,\ldots,\mathrm{K}_c) \\
\phi_{s_{K_3}} & \sim & GIW\,(\beta_s)\,, & \mathrm{K}_3 = (1,\ldots,\mathrm{K}_s) \\
\phi_{g_{K_4}} & \sim & GIW\,(\beta_g)\,, & \mathrm{K}_4 = (1,\ldots,\mathrm{K}_g) \\
\pi_{t_{K_5}} & \sim & Dir\,(\lambda) \quad , & \mathrm{K}_5 = (1,\ldots,\mathbf{K}_{\text{POS Tag States}}) \\
\pi_a & \sim & Dir\,(\alpha_a) & \\
\pi_c & \sim & Dir\,(\alpha_c) & \\
\pi_s & \sim & Dir\,(\alpha_s) & \\
\pi_g & \sim & Dir\,(\alpha_g) & \\
p_i & \sim & Cat\,(\delta) & \\
m_i & \sim & Cat\,(\pi_{\mathcal{Z}_i^t}) & \\
\omega_i & \sim & Cat\,(\theta_{m,z}) & \\
\mathcal{Z}_p^a & \sim & Cat\,(\pi_a) & \\
\mathcal{Z}_p^c & \sim & Cat\,(\pi_c) & \\
\mathcal{Z}_p^s & \sim & Cat\,(\pi_s) & \\
\mathcal{Z}_p^g & \sim & Cat\,(\pi_g) & \\
a_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^a}) & \\
c_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^c}) & \\
s_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^s}) & \\
g_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^g}) &
\end{cases}
\tag{5}
$$

The parameter K_5 represents the number of tag states (Section IV). The latent variables of the Bayesian learning model are inferred using the Gibbs sampling algorithm [16] so as to allow the model to learn correspondences between grammatical categories of words and visual cues. This algorithm repeatedly generates samples from the posterior distributions of the model parameters expressed as follows:

$$
\begin{cases}
\phi_a & \sim & \mathbb{P}\,(\phi_a \,|\, a_p\,,\beta_a) \\
\phi_c & \sim & \mathbb{P}\,(\phi_c \,|\, c_p\,,\beta_c) \\
\phi_s & \sim & \mathbb{P}\,(\phi_s \,|\, s_p\,,\beta_s) \\
\phi_g & \sim & \mathbb{P}\,(\phi_g \,|\, g_p\,,\beta_g) \\
\pi_t & \sim & \mathbb{P}\,(\pi_t \,|\, m\,,\mathcal{Z}^t\,,\lambda) \\
\pi_a & \sim & \mathbb{P}\,(\pi_a \,|\, \mathcal{Z}_p^a\,,\alpha_a) \\
\pi_c & \sim & \mathbb{P}\,(\pi_c \,|\, \mathcal{Z}_p^c\,,\alpha_c) \\
\pi_s & \sim & \mathbb{P}\,(\pi_s \,|\, \mathcal{Z}_p^s\,,\alpha_s) \\
\pi_g & \sim & \mathbb{P}\,(\pi_g \,|\, \mathcal{Z}_p^g\,,\alpha_g) \\
\mathcal{Z}_p^a & \sim & \mathbb{P}\,(\mathcal{Z}_p^a \,|\, a_p\,,\pi_a\,,w) \\
\mathcal{Z}_p^c & \sim & \mathbb{P}\,(\mathcal{Z}_p^c \,|\, c_p\,,\pi_c\,,w) \\
\mathcal{Z}_p^s & \sim & \mathbb{P}\,(\mathcal{Z}_p^s \,|\, s_p\,,\pi_s\,,w) \\
\mathcal{Z}_p^g & \sim & \mathbb{P}\,(\mathcal{Z}_p^g \,|\, g_p\,,\pi_g\,,w) \\
\theta_{m,z} & \sim & \mathbb{P}\,(\theta_{m,z} \,|\, W\,,m\,,p\,,\mathcal{Z}_p^a\,,\mathcal{Z}_p^c\,,\mathcal{Z}_p^s\,,\mathcal{Z}_p^g\,,\gamma) \\
p_i & \sim & \mathbb{P}\,(p_i \,|\, \omega_i\,,\theta_{m,z}\,,m_i\,,\mathcal{Z}_p^a\,,\mathcal{Z}_p^c\,,\mathcal{Z}_p^s\,,\mathcal{Z}_p^g\,,\delta) \\
m_i & \sim & \mathbb{P}\,(m_i \,|\, \omega_i\,,\mathcal{Z}_i^t\,,\theta_{m,z}\,,p_i\,,\mathcal{Z}_p^a\,,\mathcal{Z}_p^c\,,\mathcal{Z}_p^s\,,\mathcal{Z}_p^g\,,\pi_t)
\end{cases}
\tag{6}
$$

VI. EXPERIMENTAL SETUP

Both of a human tutor and Toyota HSR robot[4] are interacting in front of a tabletop (the major landmark). We employ 5 different objects (the attribute '***Object***' represents the object geometry modality in the learning phase): {CUP, BOTTLE, BOX[5], BALL, and TOY} of 5 different colors[6]: {WHITE, YELLOW, RED, GREEN, and BLUE} as referents and landmarks. Additionally, we employ 5 spatial prepositions: {INSIDE, BEHIND, ABOVE, NEAR, and BESIDE} in order to represent *spatial layouts and relationships* between referents and landmarks (the attribute '***Preposition***' represents the spatial layout modality in the learning phase) (Figure 4). Over and above, the robot executes 5 different actions on the objects during interaction: {HOLD, PUSH, PULL, RAISE, and PUT} (robot, object)[7].

In the cross-situational learning phase [38], the human tutor employs 60 different sentences to make the robot learn different spatial configurations of referents and landmarks. The experimental scenario is described as follows:

- The human tutor trains the robot on different spatial configurations of objects lying in a tabletop using visual information and the corresponding descriptive sentences to the different scenes.
- For every observed scene in the training phase, the visual perception systems determine the dynamics of the human arm while doing actions on objects, color and geometric characteristics of objects (segmented into point clouds), in addition to measuring the relative spatial relationships between the centroid coordinates of every object's point cloud to those of the other objects and the tabletop (Section III-B).
- For every scene-descriptive sentence, the POS tagging system defines numerical tags that represent syntactic categories of words (e.g., "(Push, **1**) (the, **5**) (Red, **2**) (Bottle, **4**) (Near, **6**) (the, **5**) (Cup, **4**)") (Section IV).
- The robot uses the probabilistic learning model to learn the meaning of the numerical tags of action verbs, spatial concepts, and object characteristics (Section V).
- In the test phase, the human tutor uses 30 sentences describing different scenes to validate the robustness of the learning phase.

VII. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed framework was trained offline on 60 sentences describing different spatial layouts of referents and landmarks of different colors in a tabletop scene, and was evaluated on other 30 sentences describing different scenes. The system was evaluated on its ability to estimate the modality of each word in a test sentence to be: *Action*, *Color*, *Preposition*, *Object*, or *Others*, and to estimate the referent and landmark referring words so as to define the direction of spatial relationship (i.e., Object A $\leftrightarrows$ Object B).

[4]The Human Support Robot (HSR) is developed by Toyota for assisting people in their daily life activities. It has a full-motion light weight body with a total of 11 degrees of freedom. The robot is equipped with stereo, Asus Xtion, and wide-angle cameras, a display screen, in addition to an array of sensors including a force-torque sensor, a laser range sensor, and an IMU sensor. The robot has one arm with a gripper that allows it to grasp objects at different heights efficiently [Toyota HSR Robot Website].

[5]In this study, the object 'BOX' is considered only as a landmark.

[6]We use color tapes to change the colors of the 5 objects.

[7]The action verbs were modeled on the robot in order to generate object-directed behaviors. The robot used the calculated distances to objects through point cloud information (Section III-B) in order to control its joins so as to execute predefined behaviors representing action verbs. Although we focus - in this study - on understanding human actions on objects, we will consider - in the future - making the robot able to generate adaptive behaviors on its own using its accumulated experience [20].

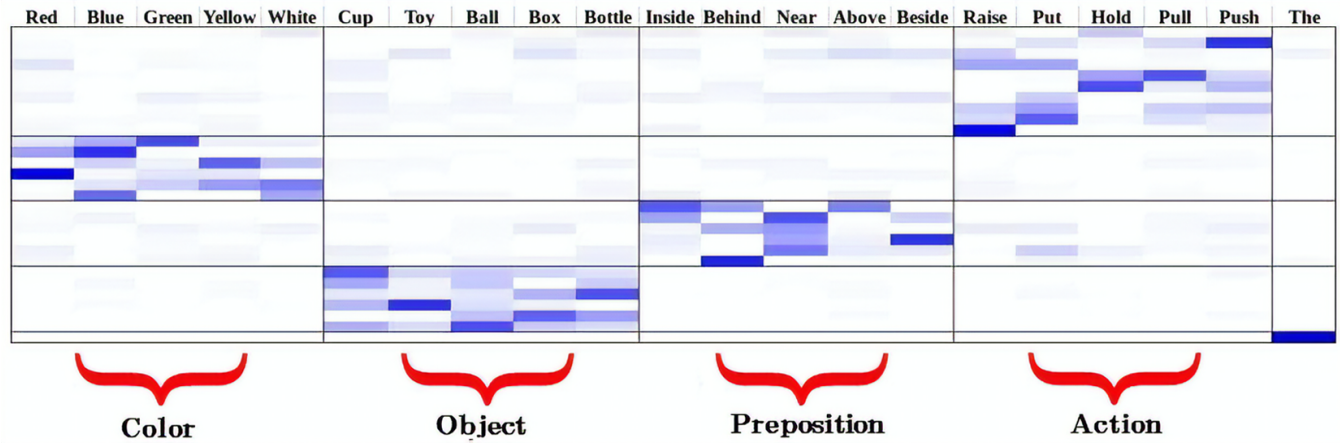


Fig. 7: Probability distribution of words over the different modalities (the dark blue color represents high probability)

TABLE II: Modality estimation for each word of the sentences of the evaluation phase.

Modality Index (/Sentences)		
Correct (Parts of Speech)		Wrong (Parts of Speech)
Action, Color, Preposition, Object	Others (The)	Color, Object
68.5%	28,6%	2,9%

While several research studies in the literature discussed language grounding through probabilistic models (e.g., Tellex et al. [47], Steels and Hild [41], Matuszek et al. [26], and Kollar et al. [22]), *and to the best of our knowledge*, no similar study addressed language grounding through a similar approach, experimental setup, or corpus to that of the current study, which makes comparing results difficult to attain. The framework successfully estimated the modality of each word in the test corpus sentences. Table (II) shows that the modalities of the different parts of speech (i.e., action, color, preposition, and object, in addition to the determiner "the") were correctly determined. This finding is clearly illustrated in Figure (7), which reveals the probability distribution of words over the different modalities. The figure shows that the patterns of data in the action, color, preposition, and object modalities are highly distinctive, among each other, and appropriately clustered. Tables (III, IV) compare the clustering results of the learning model (Gaussian-Mixture Model (GMM) clustering) and the traditional K-means method for the color and object categories. Generally, the clustering results of the model are highly similar to those of the K-means method for both the color and object categories. Except for the "Green" and "Blue" colors (where the K-means method achieved slightly higher scores), and for the "Toy" and "Cup" categories (where the GMM clustering was a bit better for the "Toy" category). The difference in the clustering performance between the GMM and K-means methods is, theoretically, related to the pattern of data itself. Consequently, the obtained results show that the representation of data in each modality was highly descriptive (Figure 7), so that both clustering methods showed a similar high performance.

TABLE III: Clustering results for the color categories with the probabilistic learning model (GMM clustering) with respect to the K-means clustering.

	Color Categories Clustering	
	Probabilistic Model (GMM Clustering)	K-means Clustering
Red	90.9%	90.9%
Green	80%	100%
Blue	45.5%	63.6%
Yellow	63.6%	63.6%
White	100%	100%

TABLE IV: Clustering results for the object categories with the probabilistic learning model (GMM clustering) with respect to the K-means clustering.

	Object Categories Clustering	
	Probabilistic Model (GMM Clustering)	K-means Clustering
Box	89.5%	89.5%
Toy	77.8%	72.2%
Cup	63.2%	78.9%
Bottle	95.2%	95.2%
Ball	100%	100%

On the other hand, the model appropriately defined the referent and landmark referring words, and the direction of their spatial relationship (i.e., Object A $\leftrightarrows$ Object B) for the different spatial prepositions, as indicated in Figure (8). These previous findings allow the robot to understand the referring words to action, color, object, and preposition, and the existing spatial relationship in a sentence to collaborate effectively with human users in space (Figure 9).

VIII. CONCLUSION AND FUTURE WORK

This research study presents a Bayesian probabilistic model for grounding action verbs, spatial concepts, and object characteristics (color and geometry) of language through visual perception during a human-robot interaction context. The proposed system grounds – through a globally unsupervised approach – the numerical tags of parts of speech based on visual cues that encode the dynamics of human actions on objects, object color and geometric

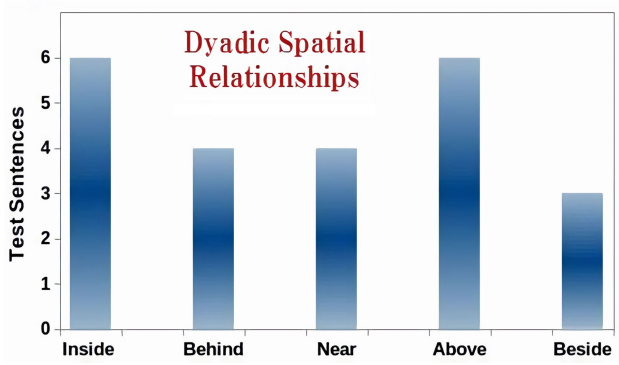
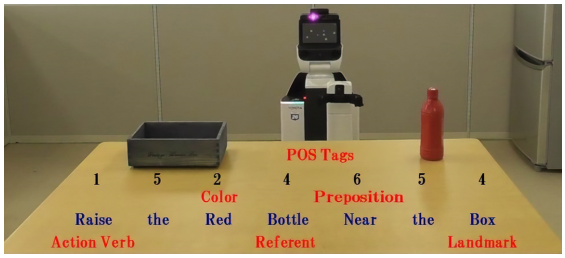
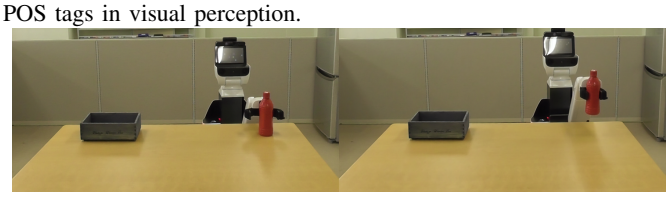


Fig. 8: Correct referent-landmark relationship estimation for the five different spatial prepositions.



(a) The robot determines the action verb, the referent and its color, and the landmark in the sentence through grounding the numerical POS tags in visual perception.



(b) The robot raises the bottle

Fig. 9: The robot successfully executes the required action on the BOTTLE located near the BOX based on visual cues.

characteristics, and spatial relationships between objects. The probabilistic framework succeeded in defining action verbs, and the referring words to objects with the colors of the referents. Additionally, the model succeeded in estimating spatial relationships between objects, so that the robot was able to perform the required actions on objects based on their point cloud information (Figure 9).

For our future work, we will extend the proposed learning model to work *online* (i.e., by online updating its learning parameters in case of new objects and spatial configurations), and to make the robot able to learn the correct referent-landmark relationships within a complex spatial configuration of objects that employ more than one landmark or referent. Additionally, we are considering to make the robot able to emulate actions on objects based on its learning experience. Over and above, we will extend the current syntactic model that could only determine syntactic categories *of* words in an unsupervised manner so as to study syntactic dependencies *between* words. This could pave the way towards addressing the induction of Combinatory Categorial Grammar (CCG) [3, 39], which accounts for syntactic and semantic representations of language, in robots through a developmentally plausible approach.

REFERENCES

[1] A. Aly, A. Taniguchi, and T. Taniguchi. A generative framework for multimodal learning of spatial concepts and object categories: An unsupervised part-of-speech tagging and 3D visual perception based approach. In *Proceedings of the 7th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*, Lisbon, Portugal, 2017. 2

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 4

[3] Y. Bisk and J. Hockenmaier. Simple robust grammar induction with combinatory categorial grammars. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, volume 2, pages 1643–1649, Toronto, Canada, 2012. 7

[4] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLC)*, Trento, Italy, 1992. 3

[5] E. Brill and M. Pop. Unsupervised learning of disambiguation rules for Part-of-Speech tagging. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech, and Language Technology*, pages 27–42. Springer, 1999. 3

[6] A. Cangelosi, K. R. Coventry, R. Rajapakse, D. Joyce, A. Bacon, L. Richards, and S. N. Newstead. Grounding language in perception: A connectionist model of spatial terms and vague quantifiers. In A. Cangelosi, G. Bugmann, and R. Borisyuk, editors, *Modeling Language, Cognition, and Action: Proceedings of the 9th Neural Computation and Psychology Workshop*, pages 47–56. World Scientific, 2005. 1

[7] C. Christodoulopoulos, S. Goldwater, and M. Steedman. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 575–584, Cambridge MA, USA, 2010. 3

[8] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLC)*, Austin TX, USA, 1988. 3

[9] P. R. Cohen, C. T. Morrison, and E. Cannon. Maps for verbs: The relation between interaction dynamics and verb use. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland, 2005. 1

[10] C. Craye, D. Filliat, and J. F. Goudou. Environment exploration for object-based visual saliency learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2016. 3

[11] C. R. Dawson, J. Wright, A. Rebguns, M. V. Escarcega, D. Fried, and P. R. Cohen. A generative probabilistic framework for learning spatial language. In *Proceedings of the 3rd Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*, Osaka, Japan, 2013. 1

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39:1–38, 1977. 2

[13] I. Farkas, T. Malik, and K. Rebrova. Grounding the meanings in sensorimotor behavior using reinforcement learning. *Frontiers in Neurorobotics*, 6(1), 2012. 1

[14] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image

analysis and automated cartography. *Communications of the ACM (CACM)*, 24(6):381–395, 1981. 3

[15] J. Gao and M. Johnson. A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 344–352, Honolulu HI, USA, 2008. 3

[16] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 6(6):721–741, 1984. 5

[17] S. Guadarrama, L. Riano, D. Golland, D. Gohring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell. Grounding spatial relations for human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2013. 1

[18] S. Harnad. The symbol grounding problem. *Physica D*, 42: 335–346, 1990. 1

[19] W. P. Headden, D. McClosky, and E. Charniak. Evaluating unsupervised part-of-speech tagging for grammar induction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, Manchester, United Kingdom, 2008. 2

[20] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura. Embodied symbol emergence based on mimesis theory. *The International Journal of Robotics Research (IJRR)*, 23:363–377, 2004. 2, 5

[21] X. Y. Jiang, U. Meier, and H. Bunke. Fast range image segmentation using high-level segmentation primitives. In *Proceedings of the 3rd IEEE International Workshop on Applications of Computer Vision (WACV)*, Sarasota Fl, USA, 1996. 3

[22] T. Kollar, S. Tellex, M. R. Walter, A. Huang, A. Bachrach, S. Hemachandra, E. Brunskill, A. G. Banerjee, D. Roy, S. J. Teller, and N. Roy. Generalized grounding graphs: A probabilistic framework for understanding grounded language. *Journal of Artificial Intelligence Research (JAIR)*, 5, 2013. 6

[23] K. Koster and M. Spann. MIR: An approach to robust clustering application to range image segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(5), 2000. 3

[24] B. Landau and R. Jackendoff. 'What' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences (BBS)*, 16:217–238, 1993. 1

[25] D. Marocco, A. Cangelosi, K. Fischer, and T. Belpaeme. Grounding action words in the sensorimotor interaction with the world: Experiments with a simulated iCub humanoid robot. *Frontiers in Neurorobotics*, 4(7), 2010. 1

[26] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, 2012. 1, 6

[27] G. Neubig. Simple, correct parallelization for blocked Gibbs sampling. Technical report, Nara Institute of Science and Technology, 2014. 4

[28] K. Ogawara, J. Takamatsu, H. Kimura, and K. Ikeuchi. Modeling manipulation interactions by Hidden Markov Models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Lausanne, Switzerland, 2002. 2

[29] P. Y. Oudeyer. *Self-Organization in the Evolution of Speech*, volume 6. Oxford University Press (OUP), Oxford, UK, 2006. 2

[30] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989. 2

[31] T. Regier. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press, Cambridge MA, USA, 1996. 1

[32] D. Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16 (3):353–385, 2002. 1

[33] D. Roy, K-Y. Hsiao, and N. Mavridis. Conversational robots: Building blocks for grounding word meanings. In *Proceedings of the International Workshop on Learning Word Meaning from Non-Linguistic Data (HLT-NAACL)*, 2003. 1

[34] R. B. Rusu, G. Bradski, and J. Hsu R. Thibaux. Fast 3D recognition and pose using the viewpoint feature histogram. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2155–2162, Taipei, Taiwan, 2010. 3

[35] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor. Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42:660–671, 2012. 1

[36] R. Schnabel, R. Wahl, and R. Klein. Efficient RANSAC for point-cloud shape detection. *Computer Graphics Forum*, 26 (2):214–226, 2007. 3

[37] J. M. Siskind. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal Of Artificial Intelligence Research (JAIR)*, 15:31–90, 2001. 1

[38] K. Smith, A. D. M. Smith, and R. A. Blythe. Cross-situational learning: An experimental study of word-learning mechanisms. *Computer Graphics Forum*, 35(3):480–498, 2011. 2, 3, 5

[39] M. Steedman. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*, 5:403–439, 1987. 7

[40] L. Steels. The symbol grounding problem has been solved, so what's next? In M. de Vega, A. Glenberg, and A. Graesser, editors, *Symbols and Embodiment: Debates on Meaning and Cognition*, pages 223–244. Oxford University Press (OUP), 2008. 2

[41] L. Steels and M. Hild, editors. *Language Grounding in Robots*, volume 1. Springer-Verlag, New York, USA, 2012. 6

[42] K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura. Learning, generation and recognition of motions by reference-point-dependent probabilistic models. *Advanced Robotics (AR)*, 25:825–848, 2011. 2

[43] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217): 1632–1634, 1995. 1

[44] J. Tani, M. Ito, and Y. Sugita. Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks*, 17(8–9), 2004. 1

[45] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh. Symbol emergence in robotics: A survey. *Advanced Robotics (AR)*, 30(11–12), 2016. 2

[46] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock. A comparative study of pose representation and dynamics modeling for online motion quality assessment. *Computer Vision and Image Understanding*, 148:136–152, 2016. 2

[47] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4):64–76, 2011. 1, 6

[48] K. Toutanova and M. Johnson. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS)*, pages 1521–1528, Vancouver, Canada, 2007. 3