Towards Understanding Syntactic Structure of Language in Human-Robot Interaction

Amir Aly[1] and Tadahiro Taniguchi[2] and Daichi Mochihashi[3]

***Abstract*— Robots are progressively moving into spaces that have been primarily shaped by human agency; they collaborate with human users in different tasks that require them to understand human language so as to behave appropriately in space. To this end, a stubborn challenge that we address in this paper is inferring the syntactic structure of language, which embraces grounding parts of speech (e.g., nouns, verbs, and prepositions) through visual perception, and induction of Combinatory Categorial Grammar (CCG) in situated human-robot interaction. This could pave the way towards making a robot able to understand the syntactic relationships between words (i.e., understand phrases), and consequently the meaning of human instructions during interaction, which is a future scope of this current study.**

I. INTRODUCTION

Creating interactive social robots able to collaborate with human users in different tasks requires high-level spatial intelligence that could make them able to discover and interact with their surroundings. Developing this spatial intelligence involves grounding language (action verbs, object characteristics (i.e., color and geometry), and spatial prepositions) and the underlying syntactic structure through sensory information so as to make a robot able to understand human instructions in the physical world.

Understanding syntactic structure of language has been intensively investigated in the literature of cognitive robotics and computational linguistics. In cognitive robotics, different research studies proposed computational models for grounding nouns, verbs, adjectives, and prepositions encoding spatial relationships between objects [1, 2, 22, 26, 38]. However, they have not investigated grammar understanding at the phrase level, which constitutes a higher level than grounding words through perception. Meanwhile, in computational linguistics, recent studies presented models for inducing combinatory syntactic structure of language [5, 15]; however, they used annotated databases for grammar induction where each word has a corresponding syntactic tag (as a noun, verb, etc.). This last point illustrates the important role that cognitive robotics could play in grammar induction through grounding parts of speech in visual perception so as to allow for learning the latent syntactic structure of phrases in a developmentally plausible manner. In this study, we build on the model of Bisk and Hockenmaier [5] for grammar induction, and propose an extended probabilistic Bayesian framework for grounding parts of speech within a cross-situational learning context between a human user and a robot [1, 32]. The overall structure of the system is coordinated through the following two phases:

1) Unsupervised calculation of Part-of-Speech (POS) tags for words (Section V), then grounding both words and tags through visual perception so as to learn their syntactic categories and meanings (Section VI)[1].
2) Unsupervised induction of Combinatory Categorial Grammar (CCG) categories *based on the grounded tags* in phase (1) (Section VIII).

The results show that the proposed probabilistic framework for grounding parts of speech was able to successfully provide correct tags to the grammar induction model. This paves the way towards grounding phrases and their induced CCG complex categories so as to allow a robot to understand phrases (not only words) composing sentences, which constitutes a direction of future research.

II. RELATED WORK

Grounding language in perception is an important challenge in artificial intelligence, cognitive robotics, and natural language processing. The "Symbol Grounding" problem was defined by Harnad [18], which discusses assigning a meaning to a meaningless symbol (e.g., new word) through interaction with the surroundings. Tanenhaus et al. [36] investigated the effect of visual cues on language understanding. Roy et al. [28] introduced an architecture to provide perceptual and affordance representations of words. Matuszek et al. [23] introduced a probabilistic framework that employs categorial grammar to develop compositional representations of language and objects in the environment. Tellex et al. [38] and Dawson et al. [11] proposed probabilistic frameworks for grounding verbs and prepositions in utterances that encode spatial relationships between referents and landmarks. Siskind [31] developed a model for grounding semantics of verbs in short image sequences. Marocco et al. [22] proposed a framework for grounding action words through sensorimotor interaction with the environment. These interesting studies, inter alia, have not discussed inferring grammatical structure of phrases in a developmentally plausible way, which constituted our motivation for the proposed study. This could open the door to make a robot understand phrases,

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

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

[3] Daichi Mochihashi is with the Institute of Statistical Mathematics, Japan daichi@ism.ac.jp.

[1]For example, the following instruction could be tagged as follows: (Raise, **1)** (the, **5)** (Red, **2)** (Bottle, **4)** (Near, **6)** (the, **5)** (Box, **4**), where these *ungrounded* numerical tags represent the syntactic categories of words (i.e., **V**erb, **Det**erminer, **Adj**ective, **Prep**osition, and **N**oun).

through an unsupervised approach, which is a future research line of this study.

The literature of natural language processing reveals different approaches towards inferring syntactic structure of language from Part-of-Speech (POS) tagging to grammar induction. Church [10], Brill [6], and Goldwater and Griffiths [17], inter alia, discussed different approaches - supervised, semi-supervised, and unsupervised - for tagging a word sequence with syntactic attributes. On the way towards studying a deeper syntactic structure of language, Klein and Manning [20] proposed a generative model for learning constituency and dependency in language for unsupervised grammar induction using induced Part-of-Speech tags. Dependency parses are determined through unlabeled edges between constituents without any defined syntactic categories, but they can not detect non-local structures efficiently. Steedman [33] introduced the rich and universal lexicalized formalism: Combinatory Categorial Grammar (CCG), where each constituent is associated with a structured syntactic category that determines its relationship to adjacent constituents in a sentence. Besides, the CCG formalism could effectively interface the syntactic to semantic structures of language [34] so as to allow a robot to better interpret human instructions.

Different approaches to unsupervised grammar induction have been investigated in the literature [5, 15]; however, these approaches disregarded learning lexical information of words and used annotated corpora. In this paper, *we bridge between artificial intelligence, cognitive robotics, and natural language processing*, and propose a framework for grounding lexical information of words through visual perception so as to infer the combinatorial syntactic structure of language - in an unsupervised manner - within a situated human-robot interaction context. This could pave the way to investigate grounding *phrases* and their induced CCG categories through visual perception so as to understand the syntactic structure of phrases composing sentences, which is a future research line of this current study that was not sufficiently addressed in the related literature.

The rest of the paper is organized as follows: Section (III) describes the system architecture, Section (IV) illustrates the visual perceptual system, Sections (V and VI) describe the lexical tagging of words and the proposed grounding model, Section (VII) presents the experimental setup, Section (VIII) introduces the CCG syntactic formalism of language, Section (IX) discusses the obtained results, and Section (X) concludes the paper.

III. SYSTEM ARCHITECTURE

The proposed framework in this study is coordinated through: (1) System for visual perception: which outputs position coordinates of the human arm joints while manipulating objects, in addition to position coordinates of objects on a tabletop and their color and geometrical characteristics (Section IV), (2) Systems for syntactic structural representation of language: which represents language through syntactic tags and combinatorial categories (Sections V and VIII), and (3) Probabilistic generative model: which grounds words and their syntactic tags through visual perception (Section VI). The following sections in the paper discuss the proposed approach in detail.

IV. VISUAL PERCEPTUAL INFORMATION

A. Skeleton Tracking: Representation of Action Verbs

The left-to-right HMM-based gesture model[2] uses the tracked (x, y, z) position coordinates[3] of the human right-arm joints (Figure 1) (converted to the local coordinate system of the referent) as observations [1]. Five HMM models are used to represent five action verbs (Section VII). Each HMM model is trained[4], during the cross-situational learning phase [32], on position coordinates of the arm joints while performing an action in different ways using the Expectation-Maximization (EM) algorithm [12]. The probabilities of evaluation of the test joint coordinates, through the different trained HMM models, are used to represent actions as observations in the probabilistic generative model (Section VI).

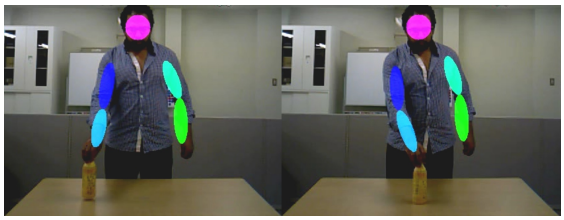


Fig. 1: Human body tracking and action characterization for object manipulation.

B. Object Segmentation into Point Cloud: Representation of Spatial Concepts

The unsupervised object segmentation model in the framework segments objects lying on a tabletop into distinct 3D point clouds with centroids representing their (x, y, z) coordinates in respect of the robot camera (Figure 2)[5]. These coordinates allow the learning model to understand spatial concepts and relationships between objects. Each point cloud is characterized using its RGB color histogram and the Viewpoint Feature Histogram (VFH) descriptor [29] that could efficiently represent object geometry and viewpoint while being invariant to scale and pose. Having calculated object locations and features, the robot employs the probabilistic

[2]Hidden Markov Models (HMM) have been intensively used in the literature to model human body motion (a time-series observation sequence) in order to enable a robot to learn and generate behaviors without temporal constraints [25, 35].

[3]The 3D tracking system uses the SDK OpenNI2 and the middleware NITE2.

[4]For each action verb in the training and test corpora (Section VII), position coordinates of the arm joints have been recorded during five different trials per action: three for training and two for testing.

[5]The model detects the tabletop plane using the RANSAC algorithm [13], and the orthogonal wall planes in contact with, at least, one image border. The remaining points in the cloud are voxelized and clustered into distinct blobs representing object candidates. This model-based approach does not require prior knowledge about the environment such as neighboring information and the number of regions to process [19, 21].

generative model (Section VI) in order to ground spatial concepts and object characteristics (i.e., color and geometry) through a cross-situational learning context with a human tutor in space [32].

V. PART-OF-SPEECH TAGGING: UNGROUNDED LEXICAL TAGS OF WORDS

The unsupervised[6] Part-of-Speech (POS) tagging (tags induction) model assigns the numerical 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) employs tags as hidden states and words as observations [14]. The probability distribution (transition) of the hidden tag states of the word sequence w is expressed as follows:

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

The emission distribution of tags over words is expressed through the probability $\mathbb{P}(\omega_i \,|\, t_i)$ of word ω_i conditioned on tag t_i. The emission and transition parameters (θ,ϕ) are characterized using multinomial distributions with Dirichlet priors $(\alpha_\theta,\alpha_\phi)$ (K stands for the number of tag states):

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

Having an unannotated corpus with a set of m sentences $\mathrm{W} = \{w_1,\ldots,w_m\}$, the model assigns the most likely tag set $\mathrm{T} = \{T_1,\ldots,T_m\}$ - inferred using the Gibbs sampling algorithm [16] - to every sentence in the untagged corpus so as to maximize the following expression:

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

In addition to the HMM-based Part-of-Speech tagging model, we examine two other common unsupervised POS tagging models - using the same corpus - and we compare between their accuracies:

- **BROWN Hierarchical Word Clustering**: which assigns each word to a single cluster using an n-gram class conditional model [7].
- **BMMM Clustering**: which assigns words to clusters using a **B**ayesian **M**ultinomial **M**ixture **M**odel [9]. This model can incorporate different additional features on both the type and token levels of words leading to a precise calculation of POS tags.

Section (IX) provides measures for the accuracies of the three taggers, and the effect of each tagger on the word grounding and grammar induction processes.

[6]The literature reports different approaches to tagging parts of speech: (1) Supervised, which employs annotated training corpora to set up tagging dictionaries indicating possible tags to words [6], (2) Semi-supervised, which employs limited annotated corpora to estimate possible tags to new word sequences [39], (3) Unsupervised, which *does not require any training corpus* to assign tags to words [8].

VI. WORD GROUNDING IN PERCEPTION: A PROBABILISTIC GENERATIVE MODEL

Figure (3) illustrates the multimodal Bayesian generative model used for grounding words (i.e., action verbs, spatial prepositions, and object characteristics) through visual perception with six observed states ω_i, $\mathcal{Z}_i^t$, a_p, c_p, s_p, and g_p. The parameters of the generative model are defined in Table (I). The state ω_i stands for each word in the sequence $\mathrm{w} = (\omega_1,\ldots,\omega_n)$, and the state $\mathcal{Z}_i^t$ stands for syntactic categories of words (Section V). The state g_p stands for the geometrical characteristics of O observed objects represented through the VFH descriptor (Section IV-B). The state s_p stands for a spatial layout between a referent and a landmark represented through their centroid coordinates (Section IV-B). The state c_p stands for the RGB color characteristics of O observed objects (Section IV-B). The state a_p stands for the arm joints locations while making actions on objects (Section IV-A). Having a spatial configuration between a referent and a landmark, the potential existing relationships between them could be expressed as follows: *Observed Objects* $O\times(O-1)$ (i.e., Referent A ⇆ Landmark B). The probabilistic distributions that characterize the Bayesian generative model are defined as follows (where *GIW* stands for a Gaussian Inverse-Wishart distribution, *Dir* stands for a Dirichlet distribution, *Cat* stands for a categorical distribution, and *Gauss* stands for a multivariate Gaussian distribution) [3]:

$$\left\{\begin{array}{lcll} \theta_{m,z_{L_1}} & \sim & Dir\,(\gamma) \;\;, & \mathrm{L}_1 = (1,\ldots,\mathrm{L}) \\ \phi_{a_{K_1}} & \sim & GIW\,(\beta_a), & \mathrm{K}_1 = (1,\ldots,\mathrm{K_a}) \\ \phi_{c_{K_2}} & \sim & GIW\,(\beta_c), & \mathrm{K}_2 = (1,\ldots,\mathrm{K_c}) \\ \phi_{s_{K_3}} & \sim & GIW\,(\beta_s), & \mathrm{K}_3 = (1,\ldots,\mathrm{K_s}) \\ \phi_{g_{K_4}} & \sim & GIW\,(\beta_g), & \mathrm{K}_4 = (1,\ldots,\mathrm{K_g}) \\ \pi_{t_{K_5}} & \sim & Dir\,(\lambda) \;\;, & \mathrm{K}_5 = (1,\ldots,\mathbf{K}_{\text{POS Tag States}}) \\ \pi_a & \sim & Dir\,(\alpha_a) & \\ \pi_c & \sim & Dir\,(\alpha_c) & \\ \pi_s & \sim & Dir\,(\alpha_s) & \\ \pi_g & \sim & Dir\,(\alpha_g) & \\ p_i & \sim & Cat\,(\delta) & \\ m_i & \sim & Cat\,(\pi_{\mathcal{Z}_i^t}) & \\ \omega_i & \sim & Cat\,(\theta_{m,z}) & \\ \mathcal{Z}_p^a & \sim & Cat\,(\pi_a) & \\ \mathcal{Z}_p^c & \sim & Cat\,(\pi_c) & \\ \mathcal{Z}_p^s & \sim & Cat\,(\pi_s) & \\ \mathcal{Z}_p^g & \sim & Cat\,(\pi_g) & \\ a_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^a}) & \\ c_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^c}) & \\ s_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^s}) & \\ g_p & \sim & Gauss\,(\phi_{\mathcal{Z}_p^g}) & \end{array}\right. \tag{4}$$

The latent variables are inferred using the Gibbs sampling algorithm [16] to allow the model to learn correspondences between words and their syntactic categories. The resulting grounded categories of words (*illustrated with an example of a tagged human-to-robot instruction in footnote (1)[1], where those numerical tags are grounded through the generative model*) are: **V**erb (representing action verbs), **Adj**ective

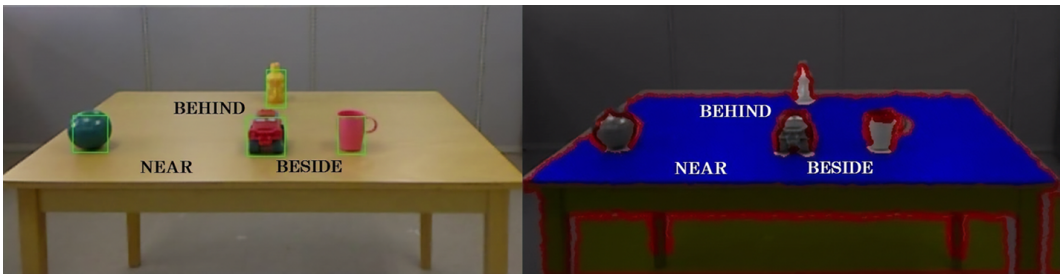


Fig. 2: Spatial concepts and relationships between objects represented through 3D point cloud information.

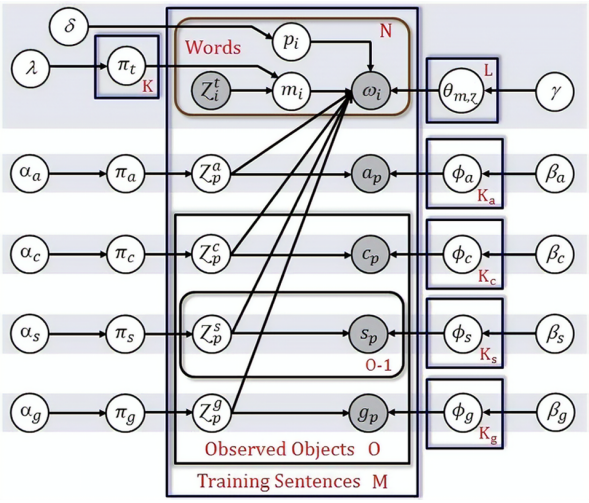


Fig. 3: Graphical representation of the probabilistic generative model. The order of words and POS tags is denoted by the index "*i*".

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

Parameter	Definition
δ	Hyperparameter of the distribution p_i
p_i	Index of spatial relationship (Object A ⇆ Object B) of each word
λ	Hyperparameter of the distribution π_t
m_i	Index of word modality ∈ {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
z_p^a	Index of action categories
z_p^c	Index of object color categories
z_p^s	Index of spatial layout categories
z_p^g	Index of object geometry categories

(representing object color), **Prep**osition (representing spatial prepositions and relationships), **N**oun (representing object geometry: object name), and **Det**erminer (representing an *others* category: the), which are used for grammar induction.

VII. EXPERIMENTAL SETUP

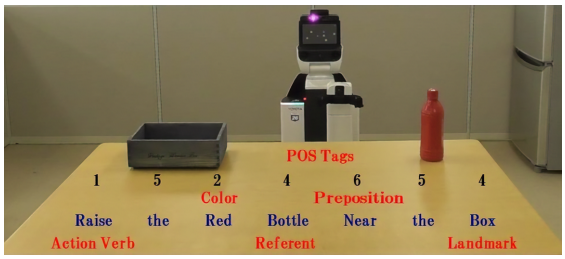
A human tutor and the HSR robot[7] (Figure 4) are interacting in front of a table on which there are five different objects: {CUP, BALL, BOTTLE, TOY, and BOX[8]} with five different colors: {GREEN, YELLOW, BLUE, RED, and WHITE} as referents and landmarks. In addition, we use five different prepositions: {ABOVE, BESIDE, NEAR, BEHIND, and INSIDE} to represent spatial relationships between objects. Moreover, the robot executes five different actions: {PUT, RAISE, HOLD, PULL, and PUSH} (robot, object)[9]. The scenario of interaction between the human tutor and the robot during the cross-situational learning phase [32] is summarized as follows:

- The tutor teaches the robot different spatial configurations of referents and landmarks lying in a tabletop - described through 60 sentences - using visual perceptual information (Section IV). The unsupervised POS tagging model calculates numerical tags representing the syntactic categories of words for every training sentence (Section V).
- The visual information describing spatial layouts characterizes the dynamics of actions, object characteristics, and spatial relationships between objects with respect to the tabletop (Section IV).
- A probabilistic model grounds words through perception in order to *define the necessary atomic categories* for unsupervised CCG categories induction (Sections VI and VIII).
- The human tutor uses 30 test sentences describing different spatial layouts of objects (characterized through similar visual perceptual information as in the training phase and numerical POS tags) in order to validate the robustness of the word grounding process (Section IX).

[7]The Human Support Robot (HSR) is developed by Toyota for providing assistance to people in daily life activities. It has a cylindrical shaped lightweight body with 11 degrees of freedom. The robot is equipped with one arm and a gripper to grasp objects, in addition to an array of sensors and cameras. [Toyota HSR Robot Website].

[8]The object 'Box' is considered only as a landmark.

[9]We focus in this study on understanding actions on objects, therefore the action verbs were modeled on the robot that used the calculated distances to object centroids to control its joins and execute predefined behaviors. As a future extension of this study, we consider making the robot able to generate actions autonomously based on it learned experience.



(a) In order to achieve the task, the robot needs to define the verb, the referent and its color, the landmark, and the existing spatial relationship between both objects through grounding words with their numerical POS tags in visual perception. The grounded POS tags are used for CCG categories induction.

(b) The robot successfully raises the red bottle near the box.

Fig. 4: The robot achieves the assigned task through grounding words and POS tags in visual perception.

VIII. COMBINATORY CATEGORIAL GRAMMAR: INFERRING SYNTACTIC STRUCTURE OF PHRASES

Combinatory Categorial Grammar (CCG) is an expressive and a lexicalized syntactic formalism [33]. Any two syntactic categories amongst the *atomic* (*S*, *N*, and *NP*), *functor* (e.g., NP/N), or *modifier* (e.g., N/N) categories of neighboring constituents could be combined through a group of rules so as to create complex categories corresponding to higher level constituents. The slash operators: "/" indicates a forward combination (e.g., an argument *follows* a functor), and "\" indicates a backward combination (e.g., an argument *precedes* a functor). The standard unary and binary combinatory rules of the CCG formalism include [4]:

1) Application combinators:

$$\overbrace{X/Y}^{\textbf{Functor}} \quad \overbrace{Y}^{\textbf{Argument}} \xrightarrow{\textbf{Forward >}} X$$

$$Y \quad X\backslash Y \xrightarrow{\textbf{Backward <}} X$$

$$(X\backslash Z)/Y \quad Y \xrightarrow{\textbf{Forward >}} X\backslash Z$$

$$Y \quad (X/Z)\backslash Y \xrightarrow{\textbf{Backward <}} X/Z$$

2) Composition combinators:

$$X/Y \quad Y/Z \xrightarrow{\textbf{Forward >B}} X/Z$$

$$Y\backslash Z \quad X\backslash Y \xrightarrow{\textbf{Backward <B}} X\backslash Z$$

3) Type-raising unary combinators (*argument* $\Rightarrow$ *functor*):

$$Y \xrightarrow{\textbf{Forward >T}} X/(X\backslash Y)$$

$$Y \xrightarrow{\textbf{Backward <T}} X\backslash(X/Y)$$

Figure (5) shows an example to illustrate the use of application combinators to create bottom-up parsing of constituents. The adjective "RED" constitutes a *modifier* category in the CCG parsing structure (i.e., it is assigned the category N/N with both arguments before and after the slash operator are identical atomic categories). This allows some parts of speech to have generic categories that modify their head arguments (e.g., adjectives *modify* nouns, and adverbs *modify* verbs), unlike a *functor* category (e.g., $X/Y \rightarrow NP/N$) that represents a head category for its dependent argument.

Grounding each word and its induced POS tag (Sections VII and V) through visual perceptual information (Section IV) using the probabilistic generative model (Section VI) produces the categories: **V**erb, **Det**erminer, **Adj**ective, **Prep**osition, and **N**oun. These syntactic categories *define the atomic categories* of the CCG formalism[10]. Having induced these atomic categories, the CCG induction model - proposed by Bisk and Hockenmaier [5] - learns the latent syntactic structure of sentences in the learning database, and generates combinatory syntactic categories for sentences in the test database (Section VII) so as to validate the robustness of the grammar induction process through comparison to a gold-standard parse structure (Section IX).

The Bayesian nonparametric HDP-CCG induction model (Figure 5) employs Hierarchical Dirichlet Processes (HDP) [37] to generate an infinite set of CCG categories based on a stick-breaking model for each Dirichlet process [30] and multinomial distributions over arguments and combinators. The resulting induced CCG categories of constituent words and phrases could be combined together using the rules explained earlier that represent the combinatorial structure of phrases. Section (IX) provides measures for the accuracy of inducing CCG syntactic categories in different cases of tagging models.

IX. RESULTS AND DISCUSSION

The framework is evaluated through its ability to induce correct CCG categories using the grounded POS tags. In this section, we provide evaluation for the accuracies of the different sub-models in the framework:

Part-of-Speech Tagging: Table (II) illustrates different measures for evaluating the robustness of the POS tagging process: **V-Measure**[11], **VI-Measure**[12], and **Many-to-One (M-1)-Measure**[13]. The difference in the V-Measure scores of the different POS taggers show that the HMM-based model was lower evaluated than the other two models.

TABLE II: Evaluation of unsupervised POS tagging through different measures.

	M-1 Measure (%)	V-Measure (%)	VI-Measure
BROWN Clustering	100	94,9	0,24
HMM Clustering	100	88,67	0,57
BMMM Clustering	100	94,9	0,24

[10]Noun Phrase (NP) = Determiner + Noun (N).

[11]It measures **homogeneity** (i.e., *optimal case*: each cluster (separate word category) contains fewer classes of tags) and **completeness** (i.e., *optimal case*: classes of tags referring to the same cluster are equal) of clusters and classes [27].

[12]It measures the variation of information of a clustering solution, so that the more the clustering is complete (i.e., high V-Measure), the lower the VI-Measure would be [24].

[13]Mapping between clusters and tags.

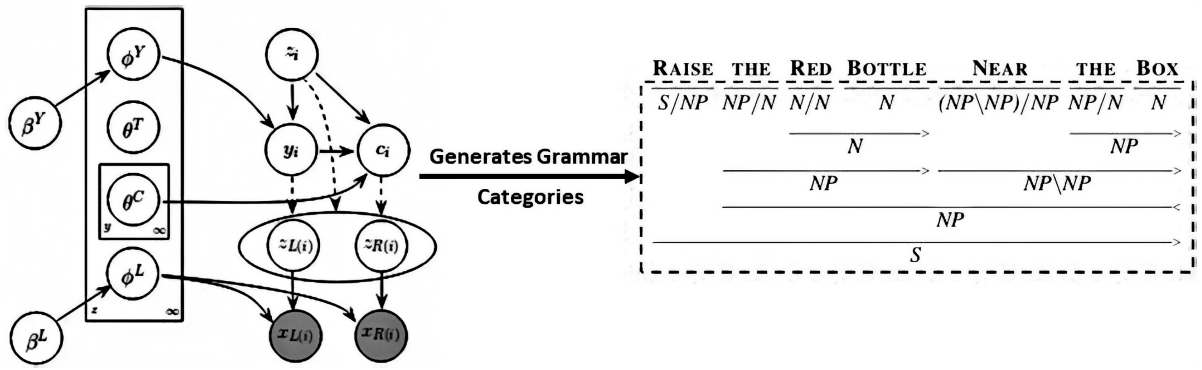


Fig. 5: Graphical representation of the CCG probabilistic generative model and an example of a resulting CCG parsing through forward and backward application combinators (maximal arity equals 2).

Having a referent and a landmark in each sentence in the corpus, the HMM-based model assigned two different tags to all referents and landmarks in the corpus (i.e., all referents had a similar tag and all landmarks had another similar tag), which could clearly reduce the completeness score of the model (Table II). Meanwhile, the BMMM and BROWN models assigned the same tag to referents and landmarks, except for some sentences in the corpus where the tags of landmarks were assigned differently, which could reduce the completeness scores of both models as well. The effect of these tagging models on the word grounding and grammar induction processes will be discussed next.

Word Grounding: Grounding words through visual perception has the objective of defining word modality and spatial relationships between objects[14]. Figure (6) illustrates the probability distribution of words over the different modalities, and shows that the patterns of data in the four modalities are highly distinctive, among each other, and appropriately clustered. Tables (III) and (IV) illustrate the results of word grounding in case of the 3 POS tagging models, which show a similar performance of the three models when estimating word modality, and indicate some differences between them when determining correct spatial relationships between referents and landmarks.

TABLE III: Estimation of word modality using different tagging models.

	Correct Word Grounding (%)			
	Verb	Adjective	Preposition	Noun (Referent & Landmark)
BROWN Clustering	76,7	100	66,7	60
HMM Clustering	73,3	100	63,3	71,7
BMMM Clustering	76,7	100	66,7	60

The HMM-based tagging model achieved higher scores, in average, than the other two models in the grounding process (more specifically in Table IV with the preposition ABOVE), despite that it had a lower V-Measure score than the other models (Table II). This could be related to that the HMM-based model was able to better provide a unified tag to all landmarks, unlike the other models as explained earlier. However, these differences do not finally affect the accuracy of the grounding process in a global scope.

[14]Despite the rich literature in language grounding, we could not find a similar study in the approach, experimental setup, or corpus to the current one, **which makes comparing these results to those of the other studies difficult to achieve**.

TABLE IV: Correct referent-landmark spatial relationships represented through different prepositions.

	Correct Spatial Relationships Between Objects (%)				
	Above	Beside	Near	Behind	Inside
BROWN Clustering	0	83,3	100	33,3	100
HMM Clustering	100	66,7	57,1	83,3	50
BMMM Clustering	0	66,7	85,7	33,3	50

CCG Categories Induction: For the CCG induction process, we use the grounded parts of speech expressed through the standard tag set of the Penn Treebank Project[15]: **V**erb: VB, **Det**erminer: DT, **Adj**ective: JJ, **Prep**osition: IN, and **N**oun: NN as input to the CCG induction model, which learns the latent syntactic structure of sentences in the learning corpus so as to generate parse trees for sentences in the test corpus. These syntactic parses are highly dependent on the grounded tags, so that wrong tags could generate imprecise parse trees. To evaluate the robustness of the CCG induction process, we use a gold-standard parse file of all sentences in the test corpus to compare against. This file contains correct POS tags and dependency relations between words in each sentence that indicate edges of standard parse trees[16]. We compare these edges to those resulting from the CCG model's predicted parses by calculating the number of matching edges.

TABLE V: Accuracy of CCG categories induction using different tagging models.

	CCG Categories Induction / Matching Edges (%)
BROWN Clustering	55,2
HMM Clustering	59,4
BMMM Clustering	60,6
Gold-POS (Without Grounding Model)	68,2

Table (V) illustrate the accuracy of CCG categories induction in case of the 3 POS tagging models in comparison to the case where the grounding model was not employed so

[15]Penn Treebank Part-of-Speech Tag Set.

[16]These syntactic dependencies between words are calculated using Stanford Parser for evaluation only.

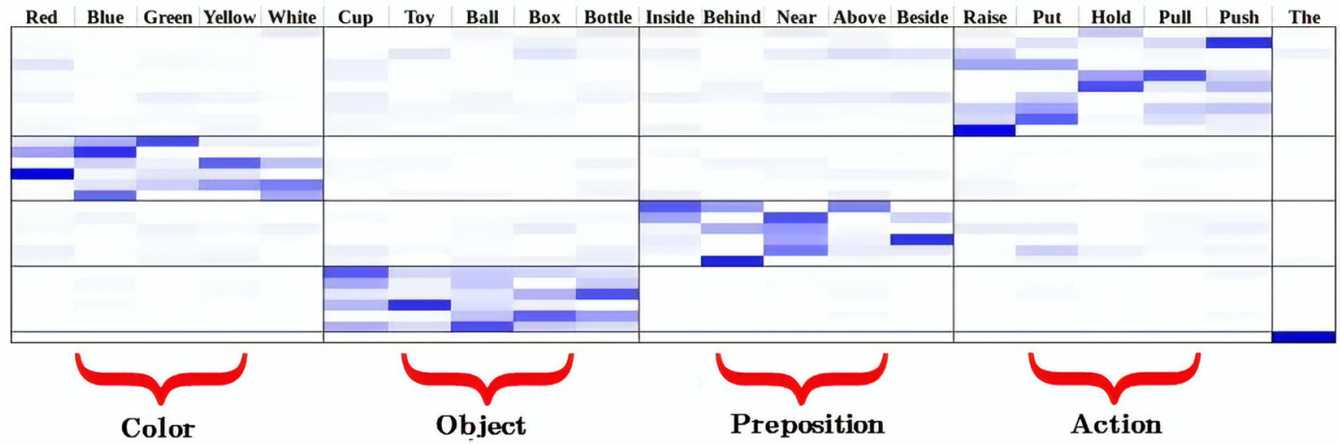


Fig. 6: Probability distribution of words over the different modalities (the dark blue color represents high probability)

that grammar induction was based only on gold-POS tags. This indicates that the BMMM and HMM-based tagging models showed, approximately, a similar good performance in grammar induction. These findings illustrate the ability of the framework to associate correct word grounding to grammar induction so as to investigate the combinatorial syntactic structure of language in an unsupervised manner. Similarly, they open the door to extend this framework to ground the generated CCG categories through perception in order to allow a robot to understand complex phrases during interaction. This includes understanding those phrases that encode spatial relationships expressed through prepositions composed of more than one word (e.g., *in front of* the Box), which would constitute a big step towards making robots able to collaborate effectively with human users in space.

X. CONCLUSION AND FUTURE WORK

This study presents a probabilistic framework for unsupervised induction of combinatory syntactic structure of language within a human-robot interaction context. The framework calculates numerical tags representing words in an unsupervised manner, and grounds them through visual perception so as to understand the syntactic categories and meaning of words. These grounded words are used for inducing CCG categories, which builds on the current state-of-the-art where a fully annotated corpus is used for grammar induction [5]. We discuss the effect of three POS tagging models on the word grounding and grammar induction processes, where the HMM-based tagging model showed a slightly better performance, in overall, than the other models. The evaluation scores of the generated CCG parses are promising and could be further improved through ameliorating the inference process of the HDP-CCG model, which we are considering to implement.

In our future work, we will consider representing the induced CCG categories of phrases in a compositional vector space with the objective of grounding phrases in perception, which would be a crucial step towards making robots able to understand language appropriately during interaction with human users.

ACKNOWLEDGMENT

This work was supported by AIP-PRISM, **J**apan **S**cience and **T**echnology Agency "**JST**". Grant number JPMJCR18Z4, Japan.

REFERENCES

[1] A. Aly and T. Taniguchi. Towards understanding object-directed actions: A generative model for grounding syntactic categories of speech through visual perception. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, 2018. 1, 2

[2] 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. 1

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 3

[4] 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. 5

[5] Y. Bisk and J. Hockenmaier. An HDP model for inducing combinatory categorial grammars. *Transactions of the Association for Computational Linguistics*, 1:75–88, 2013. 1, 2, 5, 7

[6] 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. 2, 3

[7] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4): 467–479, 1992. 3

[8] 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

[9] C. Christodoulopoulos, S. Goldwater, and M. Steedman. A Bayesian mixture model for Part-of-Speech induction using multiple features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 638–647, Edinburgh, Scotland, 2011. 3

[10] 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. 2

[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] 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. 2

[14] 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

[15] D. Garrette, C. Dyer, J. Baldridge, and N. A. Smith. A supertag-context model for weakly-supervised CCG parser learning. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL)*, pages 22–31, Beijing, China, 2015. 1, 2

[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. 3

[17] S. Goldwater and T. L. Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 744–751, Prague, Czech Republic, 2007. 2

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

[19] 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. 2

[20] D. Klein and C. D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 478–485, Barcelona, Spain, 2004. 2

[21] 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. 2

[22] 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

[23] 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

[24] M. Meila. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98 (5):873–895, 2007. 5

[25] 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

[26] O. Roesler, A. Aly, T. Taniguchi, and Y. Hayashi. A probabilistic framework for comparing syntactic and semantic grounding of synonyms through cross-situational learning. In *Proceedings of the International Workshop on Representing a Complex World: Perception, Inference, and Learning for Joint Semantic, Geometric, and Physical Understanding, in Conjunction with the IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, 2018. 1

[27] A. Rosenberg and J. Hirschberg. V-Measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, 2007. 5

[28] 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

[29] 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. 2

[30] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 5

[31] 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

[32] 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. 1, 2, 3, 4

[33] M. Steedman, editor. *The Syntactic Process*. The MIT Press, Cambridge MA, USA, 2000. 2, 5

[34] M. Steedman and J. Baldridge. Combinatory categorial grammar. In R. Borsley and K. Borjars, editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, pages 181–246. Wiley-Blackwell, 2011. 2

[35] 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

[36] 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

[37] Y-W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 5

[38] 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

[39] 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