A Bayesian Approach to Phrase Understanding through Cross-Situational Learning

Amir Aly
Emergent Systems Laboratory
Ritsumeikan University
Japan
amir.aly@em.ci.ritsumei.ac.jp

Tadahiro Taniguchi
Emergent Systems Laboratory
Ritsumeikan University
Japan
taniguchi@em.ci.ritsumei.ac.jp

Daichi Mochihashi
The Institute of Statistical Mathematics
Japan
daichi@ism.ac.jp

Abstract

In this paper, we present an unsupervised probabilistic framework to grounding words (e.g., nouns, verbs, adjectives, and prepositions) through visual perception, and we discuss grammar induction in situated human-robot interaction with the objective of making a robot able to understand the underlying syntactic structure of human instructions so as to collaborate with users in space efficiently.

1 INTRODUCTION

Different studies in the related literature of artificial intelligence discussed probabilistic frameworks for understanding the underlying syntactic structure of language. In cognitive robotics, [16, 26, 2, 1, 18] proposed computational models for grounding nouns, verbs, adjectives, and prepositions encoding spatial relationships between objects. However, these studies, inter alia, have not discovered grammar understanding at the phrase level. In computational linguistics, [4, 11] proposed models for inducing Combinatory Categorial Grammar (CCG); however, they used annotated databases (i.e., each word has a corresponding syntactic tag as a noun, verb, etc.) for grammar induction. In this study, we build on the model of Bisk and Hockenmaier [4] for categorial grammar induction, and propose an extended probabilistic Bayesian framework for unsupervised syntactic grounding of parts of speech[1] and grammar induction (*based on the grounded parts of speech and without using any annotated databases*) within a cross-situational learning context between a human user and a robot [22, 3]. 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.

2 RELATED WORK

The "Symbol Grounding" problem was defined by Harnad [14], which refers to assigning a meaning to a meaningless symbol through interaction with the environment. Tanenhaus et al. [24] discussed the effect of visual cues on language understanding. Tellex et al. [26] and Dawson et al. [7] proposed

[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 numerical tags represent the syntactic categories of words (i.e., **V**erb, **Det**erminer, **Adj**ective, **Prep**osition, and **N**oun) that would be grounded through visual perception.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

probabilistic frameworks for grounding verbs and prepositions in utterances that encode spatial relationships between referents and landmarks. Marocco et al. [16] proposed a framework for grounding action words through sensorimotor interaction with the environment. These interesting approaches, inter alia, have not discussed inferring grammatical structure of phrases, which constituted our motivation for the proposed study.

Different approaches to inferring syntactic structure of language have been investigated in the literature of computational linguistics. Church [6], Brill [5], and Goldwater and Griffiths [13], inter alia, discussed different approaches - supervised, semi-supervised, and unsupervised - for tagging parts of speech with syntactic attributes. Klein and Manning [15] proposed a generative model for learning constituency and dependency in language for unsupervised grammar induction using induced Part-of-Speech (POS) tags, but they can not detect non-local structures efficiently. Bisk and Hockenmaier [4], Garrette et al. [11] discussed probabilistic approaches to Combinatory Categorial Grammar (CCG)[2] induction; however, these approaches disregarded learning lexical information of words and used annotated corpora.

In this paper, *we bridge between cognitive robotics and computational linguistics*, and propose a generative framework for grounding lexical information of words through visual perception so as to infer the combinatorial syntactic structure of phrases within a situated human-robot interaction context. The rest of the paper is organized as follows: Section (3) describes the system architecture, Section (4) illustrates the visual perceptual system, Sections (5 and 6) describe the lexical tagging of words and the proposed grounding model, Section (7) presents the experimental setup, Section (8) introduces the CCG syntactic formalism of language, Section (9) discusses the obtained results, and Section (10) concludes the paper.

3 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 4), (2) Systems for syntactic structural representation of language: which represents language through syntactic tags and combinatorial categories (Sections 5 and 8), and (3) Probabilistic generative model: which grounds words and their syntactic tags through visual perception (Section 6). The following sections in the paper discuss the proposed approach in detail.

4 VISUAL PERCEPTUAL INFORMATION

4.1 Skeleton Tracking: Representation of Action Verbs

The left-to-right HMM-based gesture model uses the tracked 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 7). Each HMM model is trained, during the cross-situational learning phase [22], on position coordinates of the arm joints while performing an action in different ways using the Expectation-Maximization (EM) algorithm [8]. The probabilities of evaluation of the test joint coordinates, through the trained HMM models, are used to represent actions as observations in the probabilistic grounding model (Section 6).

4.2 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 coordinates in respect of the robot camera (Figure 2)[4]. These coordinates allow the learning model to understand spatial concepts and relationships between objects. Each point cloud is characterized using its RGB color histogram

[2] Steedman [23] introduced the formalism: Combinatory Categorial Grammar (CCG), where each constituent is associated with a syntactic category that determines its relationship to adjacent constituents in a sentence.

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

[4] The model detects the tabletop plane using the RANSAC algorithm [9] and the orthogonal wall planes. The remaining points in the cloud are voxelized and clustered into distinct blobs representing object candidates.

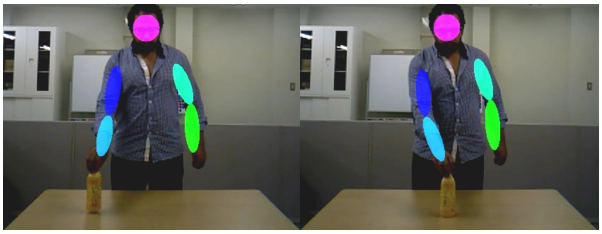

Figure 1: Human body tracking and action characterization for object manipulation.

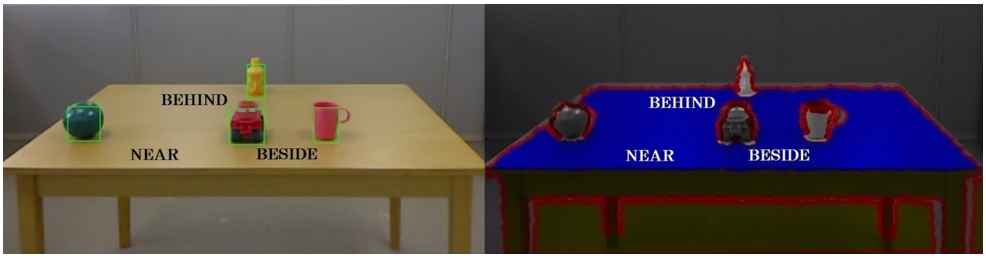

Figure 2: Spatial concepts represented through 3D point cloud information.

and the Viewpoint Feature Histogram (VFH) descriptor [20] 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 generative model (Section 6) 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 [22].

5 PART-OF-SPEECH TAGGING: UNGROUNDED TAGS OF WORDS

The unsupervised 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 [10]. 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):

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

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 $\mathrm{T} = \{\tau_1, \ldots, \tau_m\}$ - inferred using the Gibbs sampling algorithm [12] - to every sentence in the untagged corpus so as to maximize the following expression:

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

The calculated numerical tags by the HMM-based tagging model are used as observations in the probabilistic generative model (Section 6) in order to ground words, and tags, through visual perception.

6 WORD GROUNDING: A PROBABILISTIC GENERATIVE MODEL

The generative Bayesian model used for grounding words through visual perception with six observed states ω_i, $\mathcal{Z}_i^t$, a_p, c_p, s_p, and g_p is illustrated in Figure (3). The parameters of the model are defined in

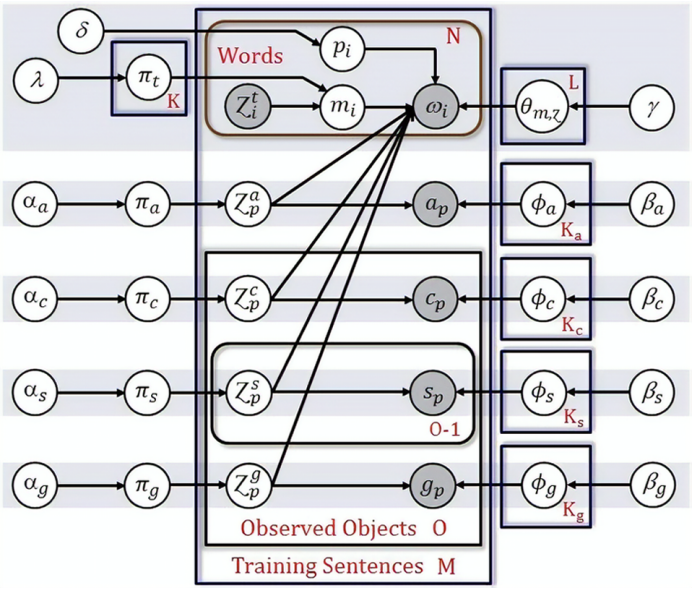

Figure 3: Graphical representation of the probabilistic generative model.

Table 1: Definition 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
$\mathcal{Z}_p^a$	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

Table (1). The state ω_i stands for each word in the sequence w = $(\omega_1, \ldots, \omega_n)$, and the state $\mathcal{Z}_i^t$ stands for syntactic categories of words (Section 5). The state g_p stands for the geometrical characteristics of O observed objects represented through the VFH descriptor (Section 4.2). The state s_p stands for a spatial layout between a referent and a landmark represented through their centroid coordinates (Section 4.2). The state c_p stands for the RGB color characteristics of O observed objects (Section 4.2). The state a_p stands for the arm joints locations while making actions on objects (Section 4.1). Having a spatial configuration between a referent and a landmark, the potential existing relationships between them could be expressed as follows: *Observed Objects* $O \times (O-1)$. The probabilistic distributions that characterize the Bayesian generative model are defined as follows (where *GIW*, *Dir*, *Cat*, and *Gauss* stand for a Gaussian Inverse-Wishart distribution, a Dirichlet distribution, a categorical distribution, and a multivariate Gaussian distribution):

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

The latent variables are inferred using the Gibbs sampling algorithm [12] to allow the model to learn correspondences between words and their syntactic categories. The resulting grounded categories of words are: **V**erb (representing action verbs), **Adj**ective (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.

7 EXPERIMENTAL SETUP

A human tutor and the HSR robot (Figure 4) are interacting in front of a table on which there are five different objects: **{**Cup, Ball, Bottle, Toy, and Box**}** 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). The scenario of interaction between the tutor and the robot 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 4). The unsupervised POS tagging model calculates numerical tags representing the syntactic categories of words for every training sentence (Section 5).
- The visual perceptual information characterizes the dynamics of actions, object characteristics, and spatial relationships between objects (Section 4).
- A probabilistic model grounds words through perception in order to *define the necessary atomic categories* for unsupervised CCG categories induction (Sections 6 and 8).
- The human tutor uses 30 test sentences describing different spatial layouts of objects in order to validate the robustness of the word grounding process (Section 9).

8 COMBINATORY CATEGORIAL GRAMMAR: INFERRING SYNTACTIC STRUCTURE OF PHRASES

Combinatory Categorial Grammar (CCG) is an expressive and a lexicalized syntactic formalism [23]. Any two syntactic categories amongst the *atomic* (*S*, *N*, and *NP*), *functor* (e.g., NP/N), or

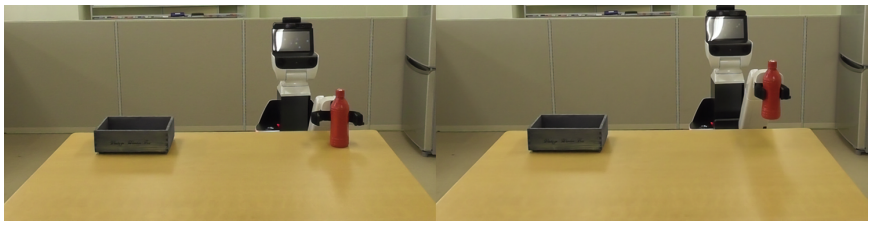

Figure 4: The robot achieves the assigned task (i.e., **RAISE THE RED BOTTLE NEAR THE BOX**) through grounding words and the calculated numerical POS tags in visual perception.

modifier (e.g., N/N) categories of neighboring constituents could be combined through a group of rules [23] so as to create complex categories corresponding to higher level constituents. The slash operators: "/" indicates forward combination (e.g., an argument *follows* a functor), and "\" indicates backward combination (e.g., an argument *precedes* a functor). The Bayesian nonparametric HDP-CCG induction model (Figure 5) employs Dirichlet Processes (DP) [25] to generate an infinite set of CCG categories defined through stick-breaking processes [21] and multinomial distributions over categories.

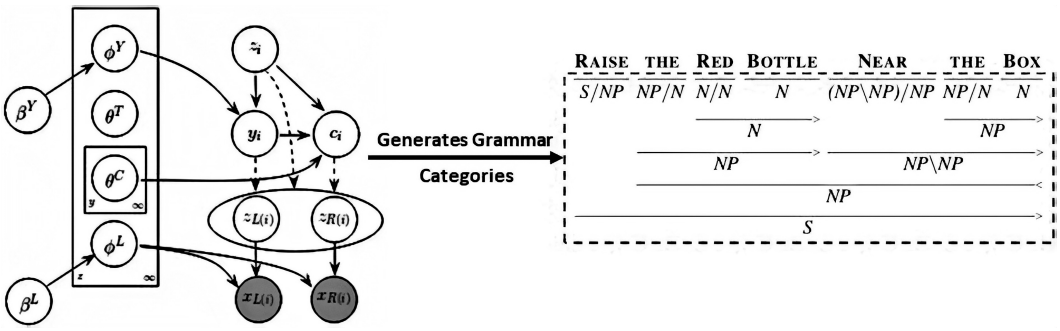

Figure 5: Graphical representation of the HDP-CCG probabilistic generative model and an example of a resulting CCG parsing through forward and backward application combinators.

Grounding each word and its induced POS tag (Sections 7 and 5) through visual perceptual information (Section 4) using the probabilistic generative model (Section 6) 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[5]. Having induced these atomic categories, the CCG induction model [4] learns the latent syntactic structure of sentences in the learning database, and generates combinatory syntactic categories for sentences in the test database (Section 7) so as to validate the robustness of the grammar induction process through comparison to a gold-standard parse structure.

9 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:

Part-of-Speech Tagging: Table (2) illustrates different measures for evaluating the robustness of the POS tagging process: **V-Measure**[6], **VI-Measure**[7], and **Many-to-One (M-1)-Measure**[8]. Having a referent and a landmark in each sentence in the corpus, the POS tagging model assigned two different tags to all referents and landmarks in the corpus (i.e., all referents had a similar tag and all landmarks

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

[6]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 [19].

[7]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 [17].

[8]It measures mapping between clusters and tags.

Table 2: Evaluation of unsupervised POS tagging through different measures.

M-1 Measure (%)	V-Measure (%)	VI-Measure
100	88,67	0,57

Table 3: Estimation of word modality grounded through visual perception.

Correct Word Grounding (%)			
Verb	**Adjective**	**Preposition**	**Noun (Referent & Landmark)**
73,3	100	63,3	71,7

had another similar tag), which could clearly reduce the completeness score (i.e., V-Measure) of the model. However, this did not affect the accuracy of the word grounding process as the model reasonably succeeded in clustering both the referents and landmarks in the "Object" category.

Word Grounding: Grounding words and POS tags through visual perception has the objective of defining word modality and spatial relationships between objects[9]. Table (3) shows that the modalities of the different parts of speech (i.e., **V**erb (Action), **Adj**ective (Color), **N**oun (Object), and **Prep**osition) were correctly determined. This finding is explained in Figure (6), which 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. Table (4) shows that the model appropriately defined the referent and landmark referring words and the direction of their spatial relationship (i.e., Referent A $\leftrightarrows$ Landmark B) for each preposition.

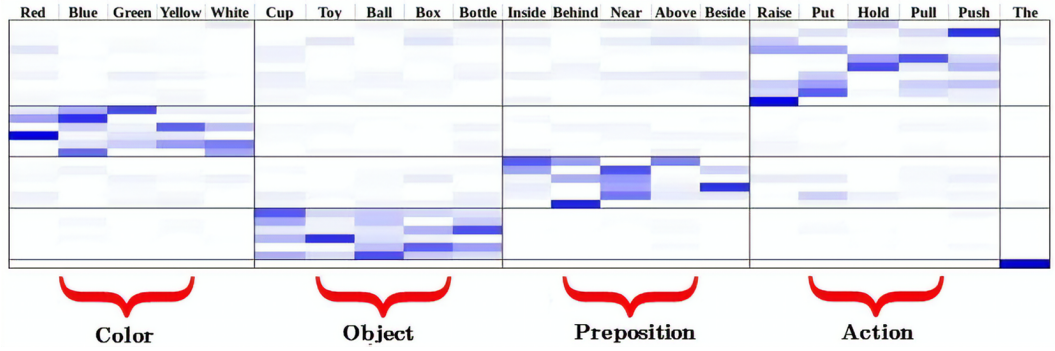

Figure 6: Probability distribution of words over the different modalities (the dark blue color represents high probability).

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[10]: **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[11]. We compare these edges to those resulting from the CCG model's predicted parses by calculating the number of matching edges.

Table (5) illustrate the accuracy of CCG categories induction in case of the grounded and gold-POS tags. It illustrates the ability of the framework to associate correct word and tag grounding to grammar

[9]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**.

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

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

Table 4: Correct referent-landmark spatial relationships represented through the different prepositions.

Correct Spatial Relationships (%)				
Above	**Beside**	**Near**	**Behind**	**Inside**
100	66,7	57,1	83,3	50

Table 5: Accuracy of CCG categories induction for the grounded and gold-POS tags.

CCG Categories Induction / Matching Edges (%)	
Grounded-POS Tags (with grounding model)	**Gold-POS Tags (without grounding model)**
59,4	68,2

induction so as to investigate the combinatorial syntactic structure of language. These findings 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.

10 CONCLUSION

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 and tags 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 [4]. The evaluation score of the generated CCG parses is promising and could be further improved through ameliorating the inference process of the HDP-CCG model, which we are considering to implement.

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] A. Aly, T. Taniguchi, and D. Mochihashi. A probabilistic approach to unsupervised induction of combinatory categorial grammar in situated human-robot interaction. In *Proceedings of the 18th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Beijing, China, 2018. 1

[4] 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, 6, 8

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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, 5

[13] 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

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

[15] 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

[16] 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, 2

[17] M. Meila. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007. 6

[18] 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

[19] 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. 6

[20] 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

[21] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 6

[22] 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

[23] M. Steedman, editor. *The Syntactic Process*. The MIT Press, Cambridge MA, USA, 2000. 2, 5, 6

[24] 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

[25] 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. 6

[26] 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