# A Probabilistic Framework for Comparing Syntactic and Semantic Grounding of Synonyms through Cross-Situational Learning

Oliver Roesler, Amir Aly, Tadahiro Taniguchi, Yoshikatsu Hayashi

# A Probabilistic Framework for Comparing Syntactic and Semantic Grounding of Synonyms through Cross-Situational Learning

Oliver Roesler[1], Amir Aly[2], Tadahiro Taniguchi[2],
Yoshikatsu Hayashi[1]

***Abstract*— Natural human-robot interaction requires robots to link words to objects and actions through grounding. Although grounding has been investigated in previous studies, none of them considered grounding of synonyms. In this paper, we try to fill this gap by introducing a Bayesian learning model for grounding synonymous object and action names using cross-situational learning. Three different word representations are employed with the probabilistic model and evaluated according to their grounding performance. Words are grounded through geometric characteristics of objects and kinematic features of the robot joints during action execution. An interaction experiment between a human tutor and HSR robot is used to evaluate the proposed model. The results show that representing words by syntactic and/or semantic information achieves worse grounding results than representing them by unique numbers.**

## I. Introduction

In 2016, 67,000 service robots have been sold worldwide resulting in a slowly growing number of human-robot interactions in peoples everyday life. To enable robots to efficiently collaborate with human users in complex environments, they must be able to converse in natural language and understand the instructions of a user so that they execute the desired actions appropriately, such as *pick up a drink* or *grab a box* [17, 24]. To achieve this goal, robots have to do "Symbol Grounding", which was first described in Harnad [15], to relate words and sensory data that refer to the same object or action to each other. Although, previous studies investigated the use of cross-situational learning for grounding of objects [11, 27] as well as spatial concepts [2, 9, 28], they ensured that one word appears several times together with the same perceptual feature vector to allow the creation of a corresponding mapping [12]. However, natural language is ambiguous due to homonymy, i.e. one word refers to several objects or actions, and synonymy, i.e. one object or action can be referred to by several different words. The latter does not need to be actual synonyms, especially, considering that according to the "Principle of Contrast" no two words refer to the exact same meaning, i.e. there are no true synonyms [6]. Consequently, words are only synonyms as references to an object or action in a particular set of situations. Examples are words that refer to the purpose or content of an object, instead of the object itself, such as: *tea* or *coffee* instead of *cup*.

[1]O. Roesler and Y. Hayashi are with Biomedical Engineering, School of Biological Sciences, University of Reading, UK (email: oliver@roesler.co.uk, y.hayashi@reading.ac.uk).

[2]A. Aly and T. Taniguchi are with the Emergent Systems Laboratory, Ritsumeikan University, Japan (email: amir.aly@em.ci.ritsumei.ac.jp, taniguchi@em.ci.ritsumei.ac.jp).

In this paper, we take a step towards grounding synonyms through a developmentally plausible approach so as to infer the meaning of objects and actions. More specifically, we present an unsupervised learning model for sensory-motor coupling using a probabilistic learning model and a robot. In the model, words are represented in three different ways: as indices, Part-of-Speech tags, or syntactic-semantic vectors, thereby, allowing the investigation of the influence of syntactic and semantic information on grounding of synonyms. The rest of this paper is structured as follows: Section (II) provides an overview of the framework. The experimental design and the obtained results are described in Sections (III and IV). Finally, Section (V) concludes the paper.

## II. System Overview

The used grounding system consists of five parts: (1) Neural Network Language Model (Word2Vec), which creates a vector space in which the distance between two vectors represents their syntactic and semantic similarity, (2) Part-of-Speech (POS) tagging system, which grammatically tags words in an unsupervised manner (i.e., *it assigns numerical tags to words without using any pre-tagged corpus or tagging dictionary*), (3) 3D object segmentation system, which determines the geometric characteristics of objects by segmenting them into point clouds, (4) Action recording system, which creates action feature vectors by recording the state of several joints while the robot is executing actions, and (5) Multimodal probabilistic learning model, which grounds object and action names through visual perception and proprioception. The inputs and outputs of the individual parts are highlighted in Figure (1), and described in detail in the following subsections.

### A. Syntactic-Semantic Representation of Words

Neural Network Language Models (NNLM) represent words as high-dimensional real-valued vectors. Several different NNLM architectures are described in the literature [3, 18, 26]. One of the main advantages of these models is the level of generalization, which is not possible to attain with simple n-gram models [21]. Word2Vec, a recently developed NNLM, uses a 2-layer neural network to create word embeddings, i.e. a vector space, for a given text corpus. Syntactically and semantically similar words are located close together [19, 20]. A corpus of 100MB of Wikipedia articles was used to train Word2Vec[1]. Several names used

[1]The corpus can be downloaded at http://mattmahoney.net/dc/text8.zip.

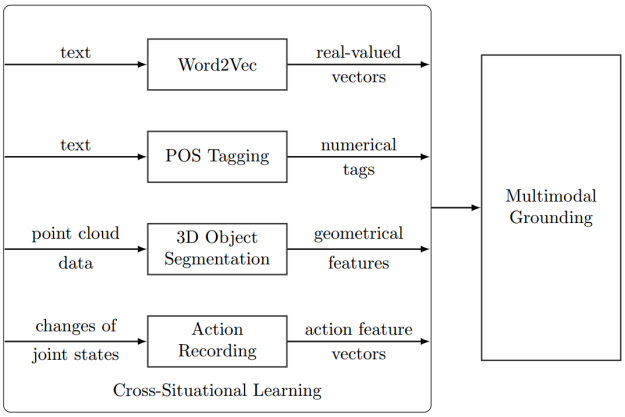


Fig. 1: Overview of the language grounding system.

TABLE I: Overview of the objects with their corresponding names.

| Object | Names | | | | |
|---|---|---|---|---|---|
| Bottle | coca_cola | soda | pepsi | coke | lemonade |
| Cup | latte | milk | milk_tea | coffee | espresso |
| Box | candy | chocolate | confection | sweets | dark_chocolate |
| Car | audi | toyota | mercedes | bmw | honda |
| Book | harry_potter | the_godfather | narnia | lord_of_the_rings | the_hobbit |

in this study are bigrams, i.e. they consist of two words, which would lead to two separate word vectors. Therefore, an underscore has been inserted between the two words to convert the original bigrams into unigrams as shown in Tables (I and II).
Multi-Dimensional Scaling (MDS) has been used to transform the vector space generated by Word2Vec to an Euclidean space [7]. Afterwards, the high vector dimensionality (100 dimensions) has been reduced through PCA to 7 dimensions so as to efficiently ground vectors in perception.

### B. Unsupervised Part-of-Speech Tagging

Through Part-of-Speech (POS) tagging words in sentences are marked with grammatical attributes (e.g., noun, verb, adjective, etc.). The literature reveals a variety of supervised, semi-supervised, and unsupervised POS tagging approaches [4, 5, 29]. In this study, an unsupervised POS tagging approach is used that induces grammatical tags for word sequences through a first-order Bayesian Hidden Markov Model (HMM), i.e., *without using any pre-tagged training corpus*[2]. The POS tagging model assigns a grammatical tag $\mathrm{T} = (t_1, \ldots, t_n)$ to each word in the sequence $\mathrm{w} = (\omega_1, \ldots, \omega_n)$. The first-order Bayesian HMM uses words as observations and tags as hidden states (Figure 2) [13].

The probability distribution of tag states for the word sequence w is defined as follows:

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

[2]For example, the POS tagging system could assign these numerical tags to words of the sentence: (Push,**7**) (the,**5**) (Coffee,**9**).

TABLE II: Overview of the used actions.

| Name 1 | Name 2 | Description |
|---|---|---|
| lift_up | raise | The object will be lifted up. |
| grab | take | The object will be grabbed, but not displaced. |
| push | poke | The object will be pushed with the closed gripper i.e. it will not be grabbed. |
| pull | drag | The object will be grabbed and moved towards the robot. |
| move | shift | The object will be grabbed and moved. |

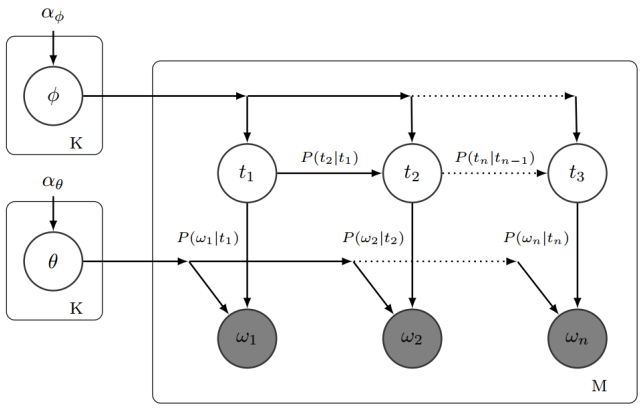


Fig. 2: Graphical representation of the HMM-based Part-of-Speech tagging model.

where the transition probability to the tag $t_i$ is conditioned on the tag $t_{i-1}$. This could encode the intuitive grammar that parts of speech might follow, like having a noun after a determiner. Emission distributions of numerical tags over words are defined through the probability $\mathbb{P}(\omega_i | t_i)$ of the word $\omega_i$ being conditioned on the tag $t_i$. For each tag state the generative transition and emission parameters of the proposed HMM model $(\phi, \theta)$ are characterized through multinomial distributions with Dirichlet priors $(\alpha_\phi, \alpha_\theta)$ (where $K$ denotes the number of tag states):

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

For an unannotated training corpus containing a set of $m$ sentences $\mathrm{W} = \{w_1, \ldots, w_m\}$, the POS tagging model tries to induce the most likely numerical tag set $\mathrm{T} = \{T_1, \ldots, T_m\}$ for each sentence in the corpus that maximizes the following expression:

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

Inferring the latent tag variables uses the Gibbs sampling algorithm [14, 22], which produces a set of samples from the posterior distribution $\mathbb{P}(T|W)$, i.e., it loops over the possible tag assignments to words that could maximize Equation (3) expressed as follows, where -$i$ denotes all samples except the $i$-th sample:

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

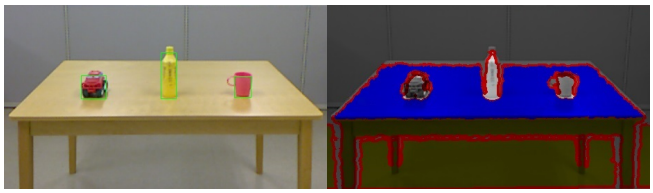

Fig. 3: Examples of the used objects and the corresponding 3D point cloud information: (A) car, (B) bottle, and (C) cup.

### C. 3D Object Features

In order to obtain object feature vectors a model based 3D point cloud segmentation approach is used due to its speed, reliability, and the fact that not much prior knowledge about the environment is required, such as object models and the number of regions to process [8, 23]. The applied model detects the major plane in the environment[3] via the RANSAC algorithm [10], and keeps track of it in consecutive frames. Planes that are orthogonal to the major plane and touch at least one border of the image are defined as wall planes, while points that are neither part of the major nor the wall planes are voxelized and clustered into blobs. Blobs of reasonable size, i.e. neither extremely small nor large, are treated as objects. Each point cloud of a segmented object is characterized through a Viewpoint Feature Histogram (VFH) [25] descriptor, which represents the geometry of the object taking into account the viewpoint and ignoring scale variance. Figure (3) shows an example of the obtained 3D point cloud information.

### D. Action Features

Action feature vectors were formulated to represent the dynamic characteristics of actions during execution through teleoperation, which could afford variations in the obtained action feature vectors. Overall, five different characteristics - each representing a possible subaction - are recorded using the sensors of the robot [30]. The employed characteristics are:

1) The distance from the actual to the lowest torso position in meters.
2) The angle of the arm in radians.
3) The angle of the wrist in radians.
4) Binary state of the gripper. (1: closing, 0: opening or no change)
5) Velocity of the base.

They are then combined into the following vector

$$\begin{pmatrix} a_1^1 & a_1^2 & a_1^3 & a_1^4 & a_1^5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_6^1 & a_6^2 & a_6^3 & a_6^4 & a_6^5 \end{pmatrix}$$

where $a^1$ represents the difference of the distances from the lowest torso position in meters, while $a^3$ and $a^4$ represent the difference in the angles of the arm and wrist in radians, respectively. The differences are calculated by subtracting the value at the beginning of the subaction from the value at the end of the subaction. $a^2$ represents the mean velocity of the base (forward/backward), and $a^5$ represents the binary gripper state. Each action is characterized through six subactions. Consequently, if an action consists of less than six subactions, rows with zeros will be added at the end.

[3]The major plane in the conducted experiment is a tabletop.

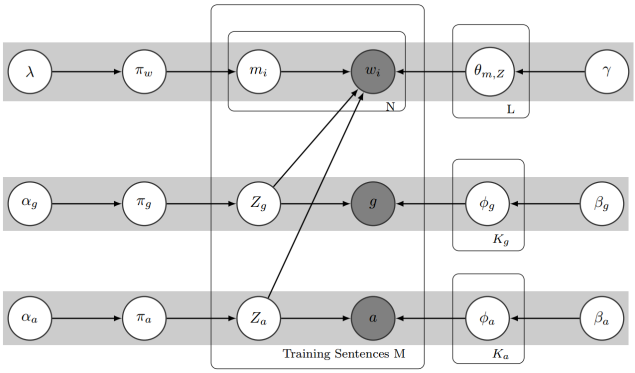


Fig. 4: Graphical representation of the probabilistic model. Indices **i**, **g** and **a** denote the order of words, object geometric features and action features, respectively.

### E. Probabilistic Learning Model

Figure (4) outlines the employed Bayesian learning model, which grounds object and action names through perception. A Bayesian network is a directed acyclic graph representing a set of probability distributions that can handle uncertainty represented by noisy perceptual data obtained from the environment [16]. Three different versions of the Bayesian learning model are employed that represent words differently: (1) Word indices, i.e. each word is represented by a unique number, e.g. (*coke*, 1) and (*lemonade*, 2), (2) POS tags, i.e. each word is represented by the grammatical category it belongs to, (3) syntactic-semantic vectors, i.e. each word is represented by a vector in a syntactic-semantic vector space. When words are represented by indices or POS tags a ***categorical*** and ***Dirichlet*** distribution are used for $w_i$ and $\theta_{m,Z}$, respectively. If words are represented by syntactic-semantic vectors a ***Gaussian*** and ***Gaussian Inverse-Wishart*** distribution are used instead.

In the probabilistic learning model, words are represented by the observed state $w_i$, which can be syntactic-semantic vectors, POS tags or indices (Sections II-A and II-B). The observed state $g$ represents the geometric characteristics of objects expressed through the VFH descriptor (Section II-C). Actions are represented by the observed state $a$ (Section II-D). Table (III) provides a summary of the definitions of the learning model parameters. The corresponding probability distributions, which characterize the different modalities in the graphical model, are defined in Equation (5), where $N$ denotes a multivariate Gaussian distribution, $GIW$ denotes a Gaussian Inverse-Wishart distribution, $Dir$ denotes a Dirichlet distribution, and $Cat$ denotes a categorical distribution. The latent variables of the Bayesian learning model are inferred using the Gibbs sampling algorithm [14].

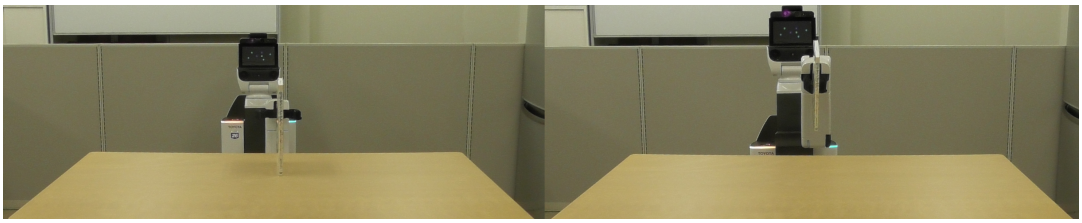

Fig. 5: Illustration of action *lift_up* executed by the robot in a tabletop scene.

TABLE III: Definitions of learning parameters in the graphical model.

| Parameter | Definition |
|---|---|
| $\lambda$ | Hyperparameter of the distribution $\pi_w$ |
| $\alpha_g$ | Hyperparameter of the distribution $\pi_g$ |
| $\alpha_a$ | Hyperparameter of the distribution $\pi_a$ |
| $m_i$ | Modality index of each word. (modality index $\in$ {Action, Object, Others}) |
| $Z_a$ | Index of action feature vector distributions |
| $Z_g$ | Index of object geometry distributions |
| $v_i$ | Word vectors |
| $g$ | Observed state representing geometric characteristics of object using VFH descriptor |
| $a$ | Observed state representing characteristics of action |
| $\gamma$ | Hyperparameter of the distribution $\theta_{m,Z}$ |
| $\beta_a$ | Hyperparameter of the distribution $\phi_a$ |
| $\beta_g$ | Hyperparameter of the distribution $\phi_g$ |

$$
\begin{cases}
w_i & \sim & Cat(\theta_{m_i,Z_{m_i}}) & \\
 & & N(\theta_{m_i,Z_{m_i}}) & \\
\theta_{m,Z_{L_1}} & \sim & Dir(\gamma) & , \quad L_1 = (1,\ldots,L) \\
 & & GIW(\gamma) & \\
\phi_{a_{K_1}} & \sim & GIW(\beta_a) & , \quad K_1 = (1,\ldots,K_a) \\
\phi_{g_{K_2}} & \sim & GIW(\beta_g) & , \quad K_2 = (1,\ldots,K_g) \\
\pi_w & \sim & Dir(\lambda) & \\
\pi_g & \sim & Dir(\alpha_g) & \\
\pi_a & \sim & Dir(\alpha_a) & \\
m_i & \sim & Cat(\pi_v) & \\
Z_g & \sim & Cat(\pi_g) & \\
Z_a & \sim & Cat(\pi_a) & \\
g & \sim & N(\phi_{Z_g}) & \\
a & \sim & N(\phi_{Z_a}) &
\end{cases}
\tag{5}
$$

## III. Experimental Setup

A human tutor and HSR robot[4] are interacting in front of a tabletop. The robot does not have any preexisting knowledge about the world, and its syntactic-semantic knowledge is limited to the word vector space and the HMM-based POS tagging model, which were created prior to training. One of the five different objects {BOTTLE, CUP, BOX, CAR, and BOOK} is placed on the table (Figures 3 and 5). Each of the objects can be referred to by five different names as shown in Table (I). During the cross-situational learning phase [11], the robot performs five different actions on each object (Figure 5), where each action can be described by two different names as illustrated in Table (II).

A total of 75 different sentences are given to the robot by the human tutor in order to allow it to ground object and action names using the recorded perceptual data. Each sentence consists of either two or three words and has one of the following two structures: "*action* the *object*" or "*action object*", respectively[5], where *action* and *object* are substituted by the corresponding names (Tables I and II). The experimental procedure consists of three phases as described below:

1) Collection of syntactic, semantic, and perceptual information for the different situations.
   a) An object is placed on the table and the robot determines its geometric characteristics so as to calculate its feature vector.
   b) A sentence is given by the human tutor to the robot, and the corresponding index, POS tag or vector of each word is obtained (Sections II-A and II-B).
   c) The human tutor teleoperates the robot to execute the given action while several kinematic characteristics are recorded and converted into an action feature vector (Section II-D).
2) The probabilistic model is used to ground words using the geometric characteristics of objects and the action feature vectors (Section II-E).
3) For the test phase, a total of 50 sentences are used to evaluate the learning framework.

In the investigated scenario, the sentences are assigned randomly to the training and test sets. Consequently, most words are used during the training and test phases, thereby, allowing the investigation whether syntactic and semantic information provides any benefit when most synonyms have already been encountered before in the cross-situational learning phase.

## IV. Results and Discussion

In several previous studies, probabilistic models have been used for language grounding [1, 9, 28]. However, *to the best of our knowledge*, none of them included synonyms and they differed in their approaches, experimental setups, or corpora from the current study, *which makes the comparison of results between our study and these studies, among many others in the literature, difficult to attain*. 20 fold cross-validation has been used, i.e. 20 different training and test

[4]The Human Support Robot from Toyota is used for the experiment. It has a cylindrical shaped body, which can move omnidirectional, and is equipped with one arm and a gripper to grasp objects. The robot has 11 degrees of freedom and is equipped with stereo and wide-angle cameras, a microphone, a display screen, and a variety of different sensors. [Official Toyota HSR Website]

[5]The latter is only used for sentences with the BOOK object. For example: "LIFT_UP HARRY_POTTER" represents the structure "*action object*", while "LIFT_UP the LEMONADE" represents the structure "*action* the *object*".

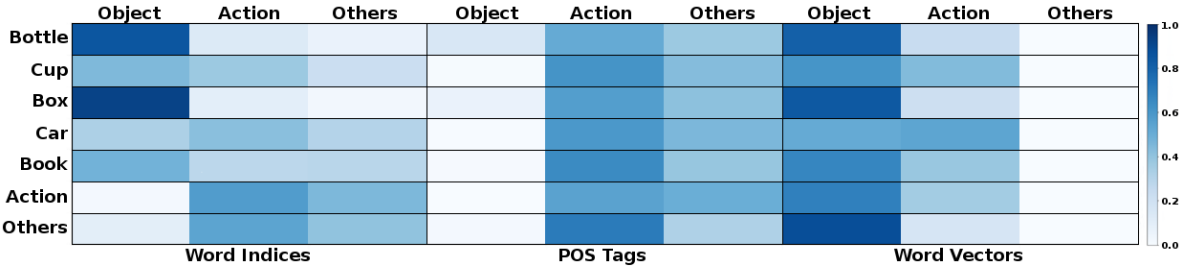


Fig. 6: Probability distributions of word categories over the different modalities for all word representations, where dark blue represents high probability.

TABLE IV: Accuracies for all modalities.

| | Object | Action | Others |
|---|---|---|---|
| **Word Indices** | 76.71% | 55.14% | 41.76% |
| **POS Tags (Syntactic)** | 4.58% | 49.56% | 33.84% |
| **Word Vectors (Syntactic-Semantic)** | 66.19% | 33.13% | 0.00% |

sets have been created. 75 sentences have been used for training, while the remaining 50 sentences have been used for the test phase.

Three different word representations have been investigated (Section II-E). The obtained accuracies show that the **Word Indices** representation achieved the best grounding for all modalities (Table IV). For the **POS Tags** representation the model did not learn the **Object** modality, which might be caused by the different object positions in the two sentence structures, i.e. "*action* the *object*" or "*action object*". We argue that this performance is due to the relatively short sentences with less syntactic word categories, which did not allow the first-order Bayesian HMM to learn the **Object** modality with respect to the previous parts of speech. For example, in the first sentence structure, the object was preceded by an action and a determiner, while in the second sentence structure, it was directly preceded by an action. Unlike our previous studies [1, 2], where the POS tagging model achieved better results using more informative sentences containing more syntactic word categories. When the **Word Vectors** representation was used the model did not learn the **Others** modality[6] because it only contains one word (the article *the*), which is not sufficient to create an independent cluster for the learning model.

Figure (6) shows the probability distributions of all word categories over the different modalities. For the **Word Indices** representation, all object categories, except the *car* names, have been correctly assigned to the **Object** modality, while the action names and the article *the* had the highest probability for the **Action** modality. In comparison, for the **POS Tags** representation the **Action** modality achieved the highest probability for all categories, while for the **Word Vectors** representation the **Object** modality achieved the highest probability for all categories, except the *car* names, which had the highest probability for the **Action** modality. The performance of the **POS Tags** was, generally, worse than the performance of the **Word Indices** due to the short sentences of the employed corpus that could not give the model enough syntactic information to learn, as explained earlier. Similarly, **Word Vectors** achieved less results with respect to **Word Indices** due to the clustering performance for the different categories based on the encoded syntactic-semantic information in vectors. Overall, this study shows that grounding of synonyms in perception, so as to allow the robot to collaborate efficiently with human users, does not necessarily benefit from the use of syntactic and/or semantic information based on the actual experimental setup of this study. However, in our future work, we will try to enhance the syntactic and semantic representation so as to achieve better results that could allow the robot to better infer the syntactic and semantic structures of a sentence.

[6]In fact, it does not assign the *Others* modality to any word as shown in Figure (6).

## V. Conclusions and Future Work

We investigated a multimodal framework for grounding synonymous object and action names through the robot visual perception and proprioception during its interaction with a human tutor. Our Bayesian learning model was set up to learn the meaning of object and action names using geometric characteristics of objects obtained from point cloud information and kinematic features of the robot joints recorded during action execution.

Our proposed model allowed the grounding of synonyms, while also showing that representing words by simple indices, instead of syntactic tags or syntactic-semantic vectors, achieves the best grounding. In future work, we will obtain grounding results for more complex sentences containing more syntactic word categories. Furthermore, we will investigate the effect of different syntactic and semantic information as well as the combination of both on grounding.

## 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 *IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, May 2018.

[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 *IEEE International Conference on Develop-*

*ment and Learning and the International Conference on Epigenetic Robotics (ICDL-EpiRob)*, Lisbon, Portugal, September 2017.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3(6):11371155, 2003.

[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.

[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*, page 2742. Springer, 1999.

[6] E. V. Clark. The principle of contrast: A constraint on language acquisition. In *Mechanisms of Language Acquisition*, pages 1–33. Lawrence Erlbaum Associates, 1987.

[7] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2001.

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

[9] C. R. Dawson, J. Wright, A. Rebguns, M. V. Escárcega, D. Fried, and P. R. Cohen. A generative probabilistic framework for learning spatial language. In *IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, Osaka, Japan, August 2013.

[10] 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, June 1981.

[11] J. F. Fontanari, V. Tikhanoff, A. Cangelosi, R. Ilin, and L. I. Perlovsky. Cross-situational learning of object-word mapping using neural modeling fields. *Neural Networks*, 22(56):579–585, JulyAugust 2009.

[12] J. F. Fontanari, V. Tikhanoff, A. Cangelosi, and L. I. Perlovsky. A cross-situational algorithm for learning a lexicon using neural modeling fields. In *International Joint Conference on Neural Networks (IJCNN)*, Atlanta, GA, USA, June 2009.

[13] 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)*, page 344352, Honolulu HI, USA, 2008.

[14] 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, November 1984.

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

[16] M. I. Jordan. Graphical models. *Statistical Science*, 19 (1):140155, 2004.

[17] C. C. Kemp, A. Edsinger, and E. Torres-Jara. Challenges for robot manipulation in human environments. *IEEE Robotics & Automation Magazine*, 14(1):20–29, March 2007.

[18] T. Mikolov, M. Karafiat, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In *Proceedings of Interspeech*, Makuhari, Chiba, Japan, September 2010.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ArXiv e-prints*, January 2013. eprint: 1301.3781.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *ArXiv e-prints*, October 2013. eprint: 1310.4546.

[21] T. Mikolov, W. t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May 2013.

[22] G. Neubig. Simple, correct parallelization for blocked gibbs sampling. Technical report, Nara Institute of Science and Technology, November 2014.

[23] A. Nguyen and B. Le. 3D point cloud segmentation: A survey. In *6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, Manila, Philippines, November 2013. IEEE.

[24] International Federation of Robotics. World robotics 2017 - service robots, October 2017.

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

[26] H. Schwenk. Continuous space language models. *Computer Speech and Language*, 21(3):492518, 2007.

[27] A. Taniguchi, T. Taniguchi, and A. Cangelosi. Cross-situational learning with bayesian generative models for multimodal category and word learning in robots. *Frontiers in Neurorobotics*, 11, 2017.

[28] 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):6476, 2011.

[29] 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)*, page 15211528, Vancouver, Canada, 2007.

[30] *HSR Manual*. Toyota Motor Corporation, 2017.4.17 edition, April 2017.