

Comparison studies on active cross-situational object-word learning using Non-Negative Matrix Factorization and Latent Dirichlet Allocation

Yuxin Chen, Jean-Baptiste Bordes, David Filliat

▶ **To cite this version:**

Yuxin Chen, Jean-Baptiste Bordes, David Filliat. Comparison studies on active cross-situational object-word learning using Non-Negative Matrix Factorization and Latent Dirichlet Allocation. IEEE Transactions on Cognitive and Developmental Systems, 2018, 10.1109/TCDS.2017.2725304 . hal-01561168

HAL Id: hal-01561168
https://hal.science/hal-01561168

Submitted on 12 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparison studies on active cross-situational object-word learning using Non-Negative Matrix Factorization and Latent Dirichlet Allocation

Yuxin Chen, *Member, IEEE,* Jean-Baptiste Bordes, and David Filliat

***Abstract*—Future intelligent robots are expected to be able to adapt continuously to their environment. For this purpose, recognizing new objects and learning new words through interactive learning with humans is fundamental. Such setup results in ambiguous teaching data which humans have been shown to address using cross-situational learning, i.e. by analyzing common factors between multiple learning situations. Moreover, they have been shown to be more efficient when actively choosing the learning samples, e.g. which object they want to learn. Implementing such abilities on robots can be performed by latent-topic learning models such as Non-Negative Matrix Factorization or Latent Dirichlet Allocation. These cross-situational learning methods tackle referential and linguistic ambiguities, and can be associated with active learning strategies. We propose two such methods: the Maximum Reconstruction Error based Selection (MRES) and Confidence Base Exploration (CBE). We present extensive experiments using these two learning algorithms through a systematic analysis on the effects of these active learning strategies in contrast with random choice. In addition, we study the factors underlying the active learning by focusing on the use of sample repetition, one of the learning behaviors that have been shown to be important for humans.**

***Index Terms*—developmental robotics, word-referent learning, cross-situational learning, active learning, Non negative Matrix Factorization (NMF), Latent Dirichlet Association (LDA).**

I. Introduction

TODAY'S robots are mainly designed to perform specialized tasks in specific and controlled scenarios. In order to be exploited at a further extent and become autonomous agents, their ability to continuously adapt to their environment and to learn to recognize new objects will be fundamental. In the last decade, implementations of effective visual features and powerful machine learning models have led to tremendous progress in the task of object recognition. However, performances are still limited in that they heavily rely on the availability of good training data that are difficult to obtain. In contrast, two years old children show an impressive ability to learn to recognize new objects during simple everyday interactions with adults. Inspired by the children's capabilities while following the developmental robotics approach, this paper aims at developing object and name learning approaches.

In this paper, we target computational models that make it possible to match objects to words for a humanoid robot using human interaction which is similar to the one taking place between children and parents. Motivated by this objective, we will compare two topic models, Non negative Matrix Factorization (NMF) and Latent Dirichlet Association (LDA), in the framework of cross-situational word-object learning. We will also study how active learning strategies could improve performances, by allowing the learner to choose which object to learn, and compare them, in terms of behaviors, with human active learning.

Yuxin Chen and David Filliat are with the U2IS, ENSTA ParisTech, Inria FLOWERS team, Universite Paris Saclay, Palaiseau, France. e-mail: {yuxin.chen, david.filliat}@ensta-paristech.fr.

Jean Baptiste Bordes is with the Ecole Polytechnique, Universite Paris Saclay, Palaiseau, France. e-mail: jean-baptiste.bordes@polytechnique.edu.

Manuscript received xx xx, 2017; revised xx xx, 2017.

A. Cross-Situational Word-Object Learning

Word-referent learning refers to the task of labeling features or objects (the so-called referents) with words in a social context. It is still an open issue when multiple mappings from word to referent can exist; a situation known as "indeterminacy of reference" [1]. Two main directions are explored by recent studies in order to decrease the uncertainty of potential referents: cross-situational learning [2], [3] and social learning [4], [5]. In cross-situational learning, the learner receives no feedback on its performance and has to analyze the common factors between different ambiguous situations. On the contrary, social learning requires the learner to receive feedback during interaction [6].

However, both methods have in common the entailment of incremental learning. Cross-situational learning provides multiple scenes related to one specific word that the learner is supposed to make use of, in order to devise the possible referents. For interactive learning, the teacher has to provide proper cues over potential mapping for the learner.

Despite the mechanisms and strategies already proposed, ambiguity still generates obstacles for learning a word's referent, that can fall in two categories:

1) *Referential ambiguity*: This ambiguity exists when several possible correspondences exist between words and visual features. This ambiguity is greater if multiple objects are presented to a learner while only one general descriptive sentence is given by the teacher. The mapping of a set of keywords (among many) to its corresponding object (among many) is also undefined.
2) *Linguistic ambiguity*: The sentences for the description of an object or a set of objects might also contain not only keywords but other grammatical words, mood words, or speaking errors as well. The algorithm thus has to distinguish keywords from other words considered as noise in this context.

B. Interactive Learning and Active Learning

The display of intelligence and curiosity by children discovering the physical world is impressive. Developmental robotics tries to bestow robots with incremental learning abilities by taking inspirations from the mental developments of infants according to the Piaget's theory of cognitive development [7]. Our paper focuses specifically on two capabilities: interactive learning and active learning, which remain as critical means of upgrading the intelligence of a robot.

Interactive learning is an important medium promoting the applications of social and biologically inspired learning mechanisms such as emulation, mimicking, imitation, and stimulus enhancement [8], [9], to help robots take full advantage of a human teacher while acquiring new skills. Whether using guidance-based methods (eg. [10], [11]) or exploration-based approaches (eg. [12], [13], [14]), the drawback of this process is the high demand of human expertise in the interaction, and improvements should be made to facilitate the rule of interaction as well as the use of ordinary information sources whose features and traits can be automatically discovered. Therefore, the present paper focuses on scenarios in which advantage is taken from weak supervision by a human teacher in order to learn words and their corresponding visual definitions. Moreover, the experiments use data recorded from speakers who do not have any knowledge of the underlying algorithms which are used for learning.

An important factor to accelerate the autonomous learning of robots is the implementation of active learning strategies. When facing the vast space of learning, curiosity often guides the infants' exploration and sets constraints for the learning processes. Therefore, artificial curiosity [15], [16] was developed for simulating the intrinsic motivation of infants for seeking new information. In this paper, active learning will be used to improve the learning performance of an agent while it learns word-referent associations.More specifically, the learner will have to choose which objects he wants the teacher to describe, depending on its current knowledge.

C. Contributions

As extension of works presented in [17] where modifications have been made on NMF to better learn a concept and in [18] in which two computational topic models have been built by applying NMF, LDA and active learning to solve cross-situational learning tasks, this paper proposes and analyses a new active learning algorithm, Confidence Based Exploration (CBE), in addition to the previously proposed one, Maximum Reconstruction Error based Selection (MRES).

This paper also goes beyond the pure performance analysis of learning algorithms and focuses also on the behavioral level, especially the difference between the proposed algorithms and humans in terms of exploitation of sample repetition. *Indeed, Immediate sample repetition* has been shown to lower referential ambiguities in word-referent learning games for humans and thus both repeated and non-repeated words were learned better [19]. It was also shown that the order of learning trials and the temporal contiguity of certain trials would also make a difference. Furthermore, Kachergis [20] shows that humans rely on immediate sample repetition in active learning situations.

In the remainder of this paper, related works of word-referent learning will be presented in Section II. We then present our own models and experimental settings in Section III and IV. The experiments are presented in Section V and focus on incremental scenarios of interactive learning. In these, our proposed models will be challenged, from the simplest experiment to more complex ones, by increasing ambiguities and noise. Finally, Section VI discusses the relations between the behavior of our models and those of humans, while Section VII draws conclusions and outlines potential future work.

II. Related Work

This problem of word-referent learning, in a larger perspective, can be stated as language grounding [21] or symbol grounding [22], [23], [24], [25]. Much of these research studies have been conducted with the goal of grounding varied content, including personal pronouns [26], [27], colors, shapes [28], set of multiple properties [29], locations [30] and spatial relations [31], [32], [33], [34], [35], [36], [37]. A lot of them, as in this paper, focus on the relatively simple, yet primary task of learning words related to object identities or features, and concentrate on studying various algorithmic aspects that influence a system's performance with data obtained by interaction with a human teacher.

Applying word-referent learning in a human-robot interactive scenario leads to both theoretical and pragmatic problems, mainly regarding the choice of *modal feature perceptions* (often concerning vision and speech), the choice of *word-referent learning algorithms*, and the choice of *learning strategies*. While the symbol grounding problems can be categorized (by [38]) as *physical symbol grounding* [39] and *social symbol grounding* [40], the word-referent learning in this paper is assumed to refer to the former one, which is aimed at grounding symbols to real world objects by a physical agent (e.g., a robot) interacting in the real world.

Yu [41] presents a multimodal system able to ground spoken names of objects in their physical referents from vocal and vision input. For the audio part, a *natural language processing* module processes raw audio data using lexical and grammatical analysis on the utterance consisting of several spoken words (i.e., keywords such as nouns) so as to convert the continuous wave pattern into a series of recognized words by considering phonetic likelihoods and grammars. For *visual processing*, a head-mounted camera is used to get visual features (including color, shape and texture description) that are extracted as perceptual representations. These feature sets are labeled with temporally co-occurring object name candidates to form many-to-many word-meaning pairs. For learning, the problem of multimodal clustering and correspondence is finally solved by the proposed *Generative Correspondence Model*.

Mangin [42] proposes an approach based on Non-negative Matrix Factorization (NMF) for learning complex human movements applied to data recordings. The learning system associates motions perceived by a camera with sound and

word labels. The motion part encodes the skeleton position and velocity acquired from a single human dancer through a Kinect device, and the sound information is a low level representation of infant directed speech sentences taken from the Acorns Caregiver dataset [43].

Araki [44] proposes a multimodal approach (including vision, sound and haptic properties of objects) that is implemented on a real robot, and focuses on learning object concepts by using Latent Dirichlet Allocation (LDA). The multimodal data are acquired autonomously by a robot equipped with a 3D visual sensor, two arms and a small handheld observation table that serves as the platform for capturing multi-view visual images of objects. This information is complemented by a small amount of linguistic information from human users.

Noda [45] uses deep neural networks to achieve the association of cross-modal information, including image, sound and motion trajectory. The memory retrieval, behavior recognition and causality modeling experiments are tested on NAO, a small humanoid robot.

The Talking Heads [23] is another model of language acquisition among a population of agents, which consists of a visual perception system, a symbolic communication channel, and an associative memory. In this experiment, a pair of agents are chosen randomly as "speaker" and "hearer" to accomplish series of guessing games on an open-ended set of geometric figures so that a shared lexicon, as well as the perceptually grounded categorization of objects are self-organized within this population without human intervention or prior specification. Learning is based on the gradual construction of categorization trees that associate features and words.

The CELL model of Roy [46] (Cross-channel Early Lexical Learning) is a cross-situational model of word-referent learning from multimodal sensory input. It has been implemented in the experiment of grounding shape names acquired through a word acquisition model based on directly processing raw data from spontaneous infant-directed speech, which are paired with video images of single objects. The main structure of CELL is composed of speech processing, computer vision, and machine learning algorithms together with STM (short-term memory) and LTM (long-term memory) settings. STM serves as a buffer where pairs of recurrent co-occurring utterance-shape events (also known as audio-visual prototypes or AV-prototypes) are filtered; and LTM further applies a recurrence filtering by first clustering the AV-prototypes from STM and then consolidating them based on a mutual information criterion [47] as the final lexical units.

A limitation of the related work is the lack of comparison of the performances of these algorithms, due to the diversity of the input data and the large spectrum of algorithmic approaches. While several models use specific learning algorithms [46],[41],[23], other use very generic topic models such as NMF [42] and LDA [44]. The latter ones provide a sound definition of the problem (i.e., finding the hidden cause that generates a visual feature and an associated word). We therefore focus on the implementation of two cross-situational learning models, based on NMF and LDA, applied on the same dataset with the goal of evaluating their strength and weaknesses in front of referential and linguistic ambiguities.

Furthermore, as presented in the introduction, active learning is a learning strategy that can improve learning performance, but its implementation has been seldom studied in this context [48]; most models simply process a pre-recorded dataset. Moreover, its implementation is strongly linked to the associated learning algorithm. While we are not proposing new active learning approaches, we study how active learning could be implemented in association with NMF and LDA, and how its behavior relates to the behavior of humans.

III. Model presentation

In this section, the framework of the proposed learning models is described, including the presentation and pre-processing of data and the learning algorithms as well as learning strategies, much of which comes from [17], [18]. Therefore, we mainly focus on the parts that are newly proposed, while already presented parts will be more briefly described.

A. Multimodal Data Presentation

The input data, noted as a corpus V of vectors V_i ($i = 1, 2, ..., n$), present two parts: the appearance of an object and an associated sentence pronounced by a human partner (see the left half of Figure 1). The main characteristic required by the learning algorithms (see Section III-C) is that these representations are additive, i.e., that it is possible to construct an object or a set of objects representation as a sum of their individual features. We used histogram representations that separate shape, color and language information to this end.

1) Histogram Presentation: The first part of each vector contains a continuous channel for the presentation of visual features, currently containing the color (V_i^{color}) and shape (V_i^{shape}) of an object. Color is encoded by an histogram of the pixel hue (from the HSV color space) of size 80. Shape is encoded by reshaping a 30x30 pixels image of the object as a 900 dimension vector (see [17] for more details). More generic features could be used as well without modifying the model. These visual features are encoded as vectors of constant size. The second part is a binary vector of the size of the dictionary of all known words (V_i^{word}) and represents the word occurrences in the sentence. The dictionary is created incrementally, starting from an empty dictionary and adding each new word encountered in sentences at the end. Note that multiple objects of interest (e.g., in Figure 2) are represented by summing the description of each individual object, thanks to the fact that the features are histograms.

In order to conduct comprehensive analysis through series of experiments covering different cases in terms of both *referential ambiguity* and *linguistic ambiguity*, six different learning scenarios are proposed and listed in Table I. Here "**Keywords only**" (i.e., *KW*) indicates the scenario where a human tutor only speaks feature-related words (ie. nouns and adjectives) in contrast to the scenario of "**Full sentence**" (ie. *FS*) in which the speaker uses natural sentences (including articles, pronouns, verbs, etc.). Besides, **Single** (ie. *S*), **Double** (ie. *D*) and **Triple** (ie. *T*) refer to how many objects (one, two and three respectively) the teacher would present simultaneously to a learner robot.

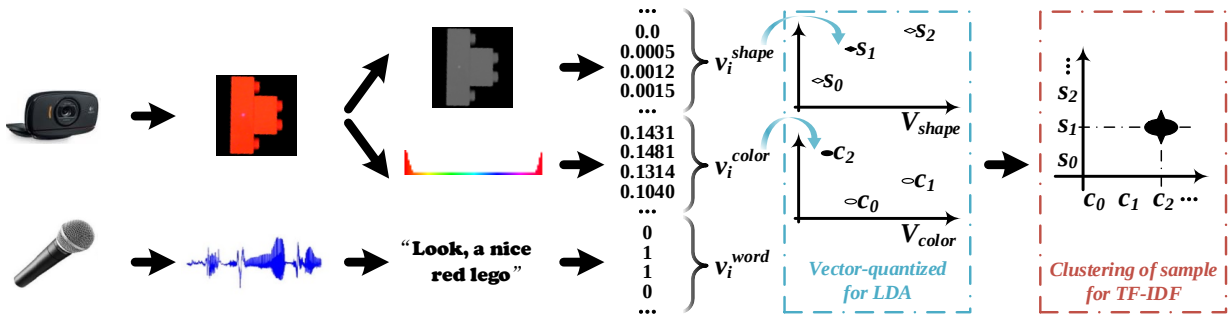


Fig. 1. Diagram of vector quantification and clustering.

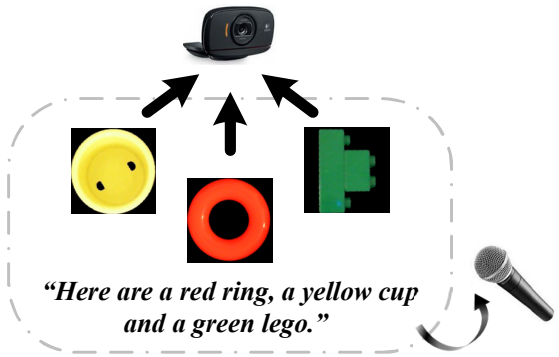


Fig. 2. Example of an ambiguous teaching situation.

TABLE I
DATA AMBIGUITIES DEFINED IN DIFFERENT SCENARIOS AND CASES.

SCENARIO \ CASE	Single	Double	Triple
Keywords only	*KW&S*	*KW&D*	*KW&T*
Full sentence	*FS&S*	*FS&D*	*FS&T*

2) Vector Quantization (VQ): In order to apply LDA and language filtering using Term Frequency-Inverse Document Frequency (TF-IDF, see below), the non symbolic (visual) channel in the observation vectors in V needs to be quantized. A simple incremental clustering is implemented along with the use of χ^2 distance, which is well adapted for histogram features:

$$\chi^2(x,y) = \sum_{k=1}^{d} (x_k - y_k)^2/(x_k + y_k)$$

This process is illustrated in the right half of Figure 1. More details can be found in [18].

Note that each of the resulting shape cluster will be labeled as $s_t \in S$, while all member vectors within a cluster will be averaged as $v_{s_t} \in V_S$, then S and V_S act as entries and corresponding contents of the shape dictionary. The same procedure takes place for the formation of the color dictionary $\{C : V_C\}$. A corpus (D) of vector-quantized samples d_i, $(i = 1, 2, ..., n)$ is then established by finding the items $s_i \in S$ and $c_i \in C$ whose member vectors are most similar to V_i^{shape} and V_i^{color}, respectively, by applying χ^2 distance. Using the words $\mathbf{w_i}$ whose corresponding indexes in V_i^{word} are positive, d_i indicates a collection of symbols, containing all words in $\mathbf{w_i}$ plus s_i and c_i.

B. Language filtering

The language filtering tries to filter out keywords from natural sentences in the *FS* scenario. LDA filters keywords thanks to its statistical properties, however, NMF provides better performance after an initial filtering of keywords [17]. It does so by relying on statistics on the word occurrences through the Term Frequency-Inverse Document Frequency (TF-IDF) popular approach [49] in text processing.

This paper inherits the modified version of TF-IDF with the use of adaptive thresholds on the IDF value, as detailed in [18]

$$\begin{aligned} idf_{low} &= idf_{min} + \eta_{low}(idf_{max} - idf_{min}) \\ idf_{high} &= idf_{min} + \eta_{high}(idf_{max} - idf_{min}) \end{aligned} \quad (1)$$

where idf_{min} and idf_{max} are the maximum and minimum of idf values for all words. The pairs (s_i,c_i) are treated as documents, while η_{low} and η_{high} values are optimized to reach the highest possible final performance on the testing set in each scenario.

C. Learning Algorithms

The learning algorithms implemented in our experiments include NMF and LDA:

1) NMF: NMF is an algorithm which computes the following decomposition:

$$\begin{aligned} V_{m\times n} &\approx W_{m\times k} H_{k\times n} \\ \begin{bmatrix} V_{shape} \\ V_{color} \\ V_{word} \end{bmatrix}_{m\times n} &\approx \begin{bmatrix} W_{shape} \\ W_{color} \\ W_{word} \end{bmatrix}_{m\times k} [H_1, H_2, ..., H_n]_{k\times n} \end{aligned} \quad (2)$$

where V is the matrix containing the observations in columns, and $W$and H are the matrices computed by NMF. W contains the k latent topics we are looking for, and H is the weights to reconstruct the observations from the topics.

W and H are solved by minimizing the Kullback-Leibler divergence:

$$D_{KL}(V \,\|\, WH) = \sum_{ij} (V_{ij} ln \frac{V_{ij}}{(WH_i)_j} - V_{ij} + (WH_i)_j) \quad (3)$$

using the multiplicative updates proposed by Lee and Seung [50].

In [17], an initialization setting of W and a normalization rule during multiplicative iterations are proposed to address the problem of instability of the decomposed results. This initialization strategy leads to an efficient concept learning in the form of "one modality-one symbol", where one word is associated to its definition in one of the feature spaces.

2) LDA: LDA is used to infer statistical correlation between visual channel and keywords. Every sample d_i is thus seen as a collection of exchangeable discrete items ω_j (which can be colors c_i, shapes s_i or words w_i) and is modeled as a generative mixture model over a set of K hidden topics $\{z_1, \ldots, z_K\}$, defined by a probability distribution on the items $p(\omega_j|z_k, \beta)$. The likelihood of a sample is thus given by (see [51] for details):

$$L_{LDA}(d_i) = \int_\theta p(\theta|\alpha) \left(\prod_j \sum_{z_k} p(z_k|\theta) p(\omega_j|z_k, \beta) \right) \quad (4)$$

where $p(\theta|\alpha)$ is a Dirichlet distribution defining the topic mixture, $p(z_k|\theta)$ is the probability of the topic z_k for this mixture and $p(\omega_j|z_k, \beta)$ the probability of an item for a given topic. The parameters to be estimated include the α of the Dirichlet distribution and the β defining the probabilities $p(\omega_j|z_k, \beta)$, which are available by maximizing the likelihood of the corpus

$$L_{LDA}(D) = \prod_{i=1}^{n} L_{LDA}(d_i)$$

using Collapsed Gibbs Sampling[1]. In practice, we observe that for a given k, the distribution $p(., z_k, \beta)$ is only significantly above zero for a couple (c_j, w_j) or a couple (s_j, w_j), thus leading to relevant word-referent associations.

D. Active learning

Our active learning strategies have the objective to select, among the available training samples describing objects, the ones that will lead to faster performance improvement. Following [15], these strategies can be either *Error Maximization based* or *Progress Maximization based*. Our investigation with Progress Maximization approaches failed to obtain satisfactory results, possibly due to the limited size of the dataset used, which is not sufficient for estimating the learning progress correctly. Besides, many existing algorithms in the literature, such as [48] ,are either unsuitable for our experimental scenario or not showing superiority over random learning performance. We therefore propose here a variant of the *Success-Threshold Strategy* and *Last Result Strategy* from [48] applied in our experiments, in addition to the previously proposed MRES [18].

1) MRES: The maximum reconstruction error based selection (MRES) favors the sample(s) with the worst reconstruction quality, as the indicator of the current limitation of the learned knowledge. Inheriting the notions from Section III-C, the generalized KL divergence is used as the reconstruction error measure with NMF:

$$D_{KL}(V_{.j} \| WH\ ) = \sum_i (V_{ij} \ln \frac{V_{ij}}{(WH_i)_j} - V_{ij} + (WH_i)_j)$$

and the likelihood function is used for LDA (eq. 4) where $V_{.j}$ represents the j_{th} sample in the training database for NMF, and d_i indicates a single document for LDA.

We then select the new training sample such that:

$$\hat{j} = \arg\max_j (D_{KL}(V_{.j} \left\| WH^{(P2)} \right.) / \textit{Gini}(H)) \quad (5)$$

for NMF and:

$$\hat{d} = \arg\min_i (L_{LDA}(d_i)) \quad (6)$$

for LDA, where $j \in \{1, 2, ..., n\}$, $d_i \in \{1, 2, ..., D\}$ are the new samples that can be considered for learning. $H^{(P2)}$ are the weight coefficients derived from H^2 by making use only of the two highest values while discarding all others, and updated via the reconstruction by minimizing $D_{KL}(V_{.j} \left\| WH^{(P2)} \right.)$. $Gini(H)$ is the Gini index that estimates the sparsity of the H vector. Note that the index $\hat{j} = \arg\max_j (D_{KL}(V_{.j} \| WH\ ))$ was used in [18], however, the version in Equation (6) produces better results, as these modifications have the objective of promoting the samples that need more than 2 dictionary components to be correctly reconstructed. This is based on the idea that these samples are not currently well known.

Equation 6 can be used to select 1, 2 or 3 samples (according to the scenario) with the worst reconstruction. However, in practice, if certain sample(s) which are chosen do not efficiently improve the current knowledge status, this/these sample(s) will be selected over and over again, resulting in a stagnation of the performance. A slack strategy introducing some stochasticity has therefore been applied and proved to efficiently improve the learning progress, especially for learning via NMF. Instead of choosing the 1, 2 and 3 samples with worst reconstruction, we randomly selected these samples among the 6, 9 or 12 ones with the worst reconstruction so as to ensure some diversity in the object choice.

2) CBE: The Confidence-based exploration (CBE) is derived from *Success-Threshold Strategy* and *Last Result Strategy* in [48], with the idea that a learner seeks to explore unknown objects only when confident enough about what has been learned already. In practice, we calculate the average reconstruction error of all used samples as the confidence indicator, noted as $D_{KL}^{n'}$ and $L_{LDA}^{D'(d)}$, for NMF and LDA, respectively:

$$\begin{aligned} D_{KL}^{n'} &= \textstyle\sum_{j=1}^{n'} [D_{KL}(V_{.j} \| WH\ )]/n' \\ L_{LDA}^{D'} &= \textstyle\sum_{d=1}^{D'} [L_{LDA}(d)]/D' \end{aligned} \quad (7)$$

where n' and D' represent number of used samples in NMF and LDA learning.

When feeling confident, $D_{KL}^{n'} \leq threshold_{D_{KL}}$ $L_{LDA}^{D'} \geq threshold_{L_{LDA}}$ the learner will explore by randomly choosing candidate objects among those that do not contain the features

[1]We use the implementation of https://github.com/ariddell/lda with all parameters initialized with default settings

[2]computed by finding the H that minimizes $D_{KL}(V_{.j} \| WH\ )$ first



Fig. 3. The 39 objects used for the experiments.

(relating color and shape) which already exist in any used samples. On the contrary, when the confidence is not sufficient, sample selection will favor already encountered color and shapes. We choose the worst reconstructed sample(s) by using the previous MRES method[3] from the pool of samples that have features already encountered in previous samples.

Note that as the incremental learning proceeds, there are less and less unknown features up to the point where all features have been seen at least once, while there still remain unused samples. In that case, the explorer has no samples to choose from, and resorts to random choice among all unused samples.

IV. Experimental Setup

In the experiment, a camera is installed over a table, facing down to capture objects while a microphone is used for receiving voice signals. The database consists in 39 objects (Figure 3), corresponding to 5 colors and 10 shapes. Therefore, 15 keywords exist in total.

A. Training stage

153 samples are recorded by ten volunteers, every object is described at least three times. Each object is described by at least two keywords as well as some others words (such as "this is", "that is"), the mean recorded sentence length is 4.026 words. A training set is established by selecting 3 samples for each one of the 39 objects (a total of 117 samples) while the remaining 36 samples, covering all the keywords, act as testing data for the evaluation of learning performance.

We perform training of the two algorithms by optimizing Equation 3 for NMF and Equation 4 for LDA. In order to focus on the performances of active learning approaches, we do not use the incremental version of NMF and LDA but simulate incremental learning by performing batch training with a growing set of samples. New samples are selected either randomly or from one of the active learning strategies.

B. Testing stage

After each training stage, the testing consists in a simulation of the situation where the learner has to find an object described by the teacher(denoted as "T2img", for "Text to Images", in the remainder of the paper). Technically, the teacher utters a textual description encoded as a binary format T_j about an object j and the learner has to choose the right object from the pool of all 36 testing objects. The testing protocol differs according to the method which is used:

[3]without the slack strategy

1) Testing with NMF:

1) The coefficient vector of hidden topics H_i associated with the visual description of each testing object i is computed by minimizing the distance $D_{KL}([V_i^{shape}, V_i^{color}] \parallel [W^{shape}, W^{color}]^T H_i)$.
2) The textual description of each object is reconstructed using the formula: $V_i^{word} = W^{word} H_i$.
3) The object in the testing set whose textual description is the closest to T_j is found by computing $\chi^2(T_j, V_i^{word})$ for all i.

2) Testing with LDA:

1) The hidden topic distribution associated to T_j is estimated using the following formula: $P(\mathbf{z}|T_j)$,
2) The associated vision feature channel is reconstructed using $P(\omega_j|T_j) = \sum_k P(\omega_j|z_k, T_j) \cdot P(z_k|T_j)$, with $\omega_j \in S \cup C$.
3) For every testing sample d_i, the log-likelihood $L(d_i|T_j) = \sum_l Cnt(\omega_l) \cdot \ln P(\omega_l|T_j)$ is computed, with $\omega_l \in S \cup C$, where $Cnt(\omega_l)$ is the number of occurrences of the visual cluster ω_l from the testing sample d_i.
4) The object whose likelihood is the highest among the candidates is taken as the output of the system.

In both cases, the percentage of right answers among the 36 testing objects is used as the performance rate.

V. Performance analysis of active learning

Previous works [17], [18] have established the following conclusions :

1) With the currently proposed experimental protocol, NMF achieves high performance for learning the objects of the database when the sentence consists only in keywords and with one object per sample.
2) NMF with TF-IDF filtering of the full sentences is able to learn with good performance in scenarios of linguistic noisy data with one object by sample.
3) Both NMF and LDA prove to be compatible with the implementation of active learning with MRES for incremental learning tasks.

In the current paper, the following experiments focus on how active learning (i.e., MRES as well as CBE) improves the learning progress in more complex scenarios. We perform automatic determination of the dictionary size (i.e., the number of topics for NMF or components for LDA) as presented in [18][4] and will allow the possibility of choosing among all the 117 training samples for the next learning step, so as to highlight the ability of active learning to efficiently ignore the already known samples.

The simulation experiments were performed 50 times in total. The results are presented by the curves of the 75^{th}, 50^{th} and 25^{th} percentile of performance among all repetitions Figure 5 and 6 illustrate results for training using *FS* data and Figure 8 and 9 in Appendix A for *KW* data. Note that in order

[4]At each training step, being K the previous number of topics, the learner will learn using both K and $K+1$ topics. If the later setting has lower error, the topic number will be increased by 1.

to enhance the comparability among all cases, we deliberately let initial samples (at the first learning step) be the same.

To facilitate quantitative comparison of the different curves, the area under the curves is computed and we use it as a global measure of learning performance. Figure 4 shows the relative performance (area under the curves, normalized by being divided by the maximum value of all cases) in different cases by using different learning models and data, which will be discussed subsequently.

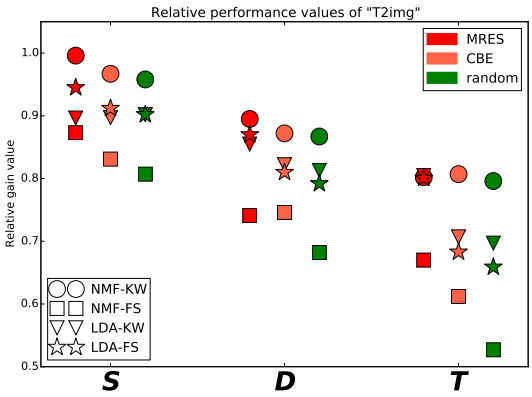


Fig. 4. Relative performance values of the models for the "T2img" image recognition task. *S*, *D* and *T* refer to how many objects (one, two and three respectively) are presented to the learner, colors indicate the sample choice strategy and shapes indicate the learning algorithm used.

A. Active learning vs. random learning by applying NMF

The performances of NMF with Full Sentences (*FS*) data is illustrated in Figure 5 (performance with *KW* data are in Figure 8). In terms of the effects of active learning, first of all, the superiority over random choice is clearly visible in the 50^{th} percentile curve in terms of speed of progress and overall quality when single object training data (*S*) are used, especially for MRES. However, this difference is limited in most cases as the 25^{th} and 75^{th} percentile curve often overlap. The difference in learning progress is however more clear in the most difficult cases (full sentences with 3 objects - FS&T). Considering the area under the curves (Figure 4), MRES exhibits overall better quality than CBE when *S* and *D* data are utilized, yet CBE performs better with *KW&T* data.

In these results, even if the active learning strategies contribute strongly to the good performance, the effect of TF-IDF should not be overlooked. Indeed the optimization of the parameters settings (see Section III-B) for one type of experiment cannot accommodate all possible sequences of learning samples. This is particularly true for the random choice strategy, where a lot of samples are used multiple times, while for MRES and CBE, we could easily find the best parameters that lead to the results illustrated in Figure 5.

B. Active learning vs. random learning by applying LDA

As shown in Figure 6 and 9, the 50^{th} percentile curve shows an improvement on the learning speed for active learning scenarios, but a clear difference is only notable in the more ambiguous situation. This difference is particularly visible in the triple scenario, in which MRES clearly outperforms CBE, which is very close to the random choice

A point to be noted is that active learning strategies achieve higher performance with LDA when linguistic noise is added compared to the case with keywords only (see discussion below).

C. Discussion regarding the overall performances

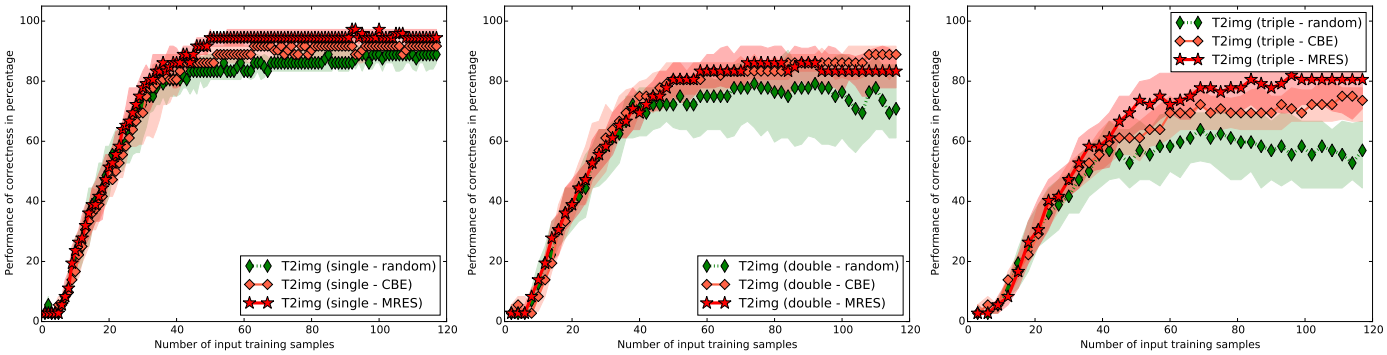
Looking at Figure 4, it makes sense that the increase on referential ambiguity (i.e., using data from *S*, *D* to *T*) leads to decrease in performance under the same scenario and with the same strategy. Besides, NMF performs better almost always than LDA in *KW* scenario. However, LDA outperforms NMF in *FS* scenario, which corresponds to the conclusion about the different abilities of dealing with linguistic ambiguities between NMF and LDA.

1) Comparing performances of KW and FS: The performances using *KW* are supposed to surpass those using *FS* as the scenario is less ambiguous in terms of linguistics. Notably, in Figure 4, exceptions can be found in the performance of LDA learning, using *S* data while applying MRES, CBE and random choice, respectively, as well as using *D* data while applying MRES.

The explanation is related to the effect of noise on the learned topics. Indeed a small part of the topics are noisy: either multiple words corresponding to a vision feature, a word associating a vacant meaning or a wrong feature. The difference between KW and FS lies in the fact that in FS, are almost always the noisy words that result in the above ill-defined topics. However, in KW this occurs for keywords. As a consequence, when performing "T2img" tests, the learner is more misled by the noisy topics in the KW case (corresponding to a real keyword) than in the FS case (corresponding to a noisy word).

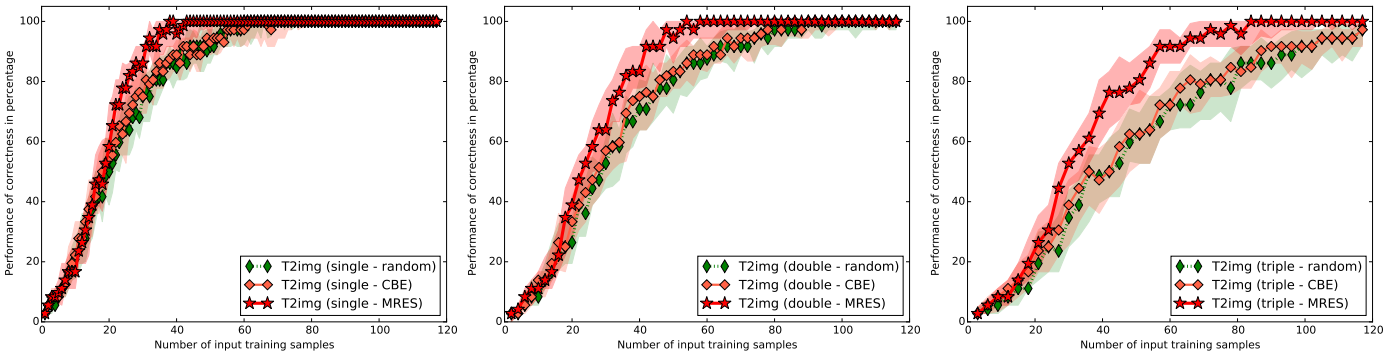
2) Comparing performances of MRES and CBE: Based on the results displayed in Section V-A and V-B, MRES's performances are better than random choice learning strategies in most scenarios, but it only shows clear superiority in the most complex scenarios. As for CBE, its gain is less clear, as the results most of the time overlap with those of random sampling. We therefore conclude that MRES is the most relevant strategy, but only in the more complex scenarios that deal with part of the complexity of human language.

In practice, the parameters have to be tuned carefully so as to exhibit as much potential of the strategies as possible. In MRES, the adopted slack strategy (Section III-D) has been observed as important in improving the performance with NMF using data of more ambiguities (ie. "D" or "T" scenarios). However, even without the slack strategy, MRES can still work efficiently with LDA, which remains effective in outperforming random choice. In CBE, finding the correct confidence threshold, whose optimized value differs according to different experimental scenarios, is essential to obtain good performances, showing a greater sensibility of this method to parameter settings. While this sensibility makes it difficult to draw definitive conclusions on the absolute algorithm perfor-



(a) "T2img" results using $FS\&S$ (b) "T2img" results using $FS\&D$ (c) "T2img" results using $FS\&T$

Fig. 5. Comparisons in performances between MRES, CBE and random choice by applying NMF with *FS* data.



(a) "T2img" results using $FS\&S$ (b) "T2img" results using $FS\&D$ (c) "T2img" results using $FS\&T$

Fig. 6. Comparisons in performances between MRES, CBE and random choice by applying LDA with *FS* data.

mances, the higher stability of LDA and MRES makes them a better choice for a more realistic future application.

VI. Behavioral analysis of active learning

In [20], comparative studies of active learning and random learning have been tested on human participants. Kachergis et al first proposed that most learners use immediate repetition to disambiguate pairings, and then further confirmed that those who repeat multiple pairs per trial outperform those who repeat only one pair per trial. Their experiment, learning names of objects presented by groups of 4, was quite different from ours, and comparing human performances on our task and theirs would make little sense. However, from the point of view of our model, learning 4 different object names or learning two different features of two objects is equivalent as these situations would be both represented by sums of histograms representing objects or features. We therefore compare the behavior of our model to the reported human behavior with the objective of testing whether the choice made by humans could be closer to the choices made by the MRES or CBE algorithms.

A. Use of instant repetitions

In order to evaluate this repetition behavior on our proposed model, we choose the learning model of LDA which appears more compatible, as previously concluded, with the implementation of active learning strategies. The cases of *KW* and *FS* with data *S*, *D* and *T* respectively, using both MRES and CBE, were chosen. We performed the protocol detailed in Section V-B and compute "R-NXT", which denotes the average number of repetition of descriptive keywords for samples in successive trial, equivalent to the term "immediate repetition" in [20]. We obtained the statistical results presented in Table II.

TABLE II
AVERAGE NUMBER OF WORD REPETITIONS IN SUCCESSIVE TRIAL AND FEATURE REPETITION WITHIN TRIAL IN ACTIVE LEARNING EXPERIMENTS USING LDA LEARNING MODEL.

Repetition / *Strategy* / Case		**MRES**		**CBE**		**random**	
		R-NXT	R-WHT	R-NXT	R-WHT	R-NXT	R-WHT
	S	0.29	*0.00*	0.31	*0.00*	0.31	*0.00*
KW	***D***	0.84	*1.00*	1.15	*0.34*	1.17	*0.30*
	T	1.74	*2.04*	2.38	*0.89*	2.41	*0.90*
	S	0.34	*0.00*	0.32	*0.00*	0.32	*0.00*
FS	***D***	0.98	*1.00*	1.16	*0.34*	1.17	*0.31*
	T	1.92	*2.02*	2.38	*0.87*	2.41	*0.88*

Compared to human behavior, the strategy resulting from our algorithms seems different: in the *T* scenario, for instance, random sample choices (for both *KW* and *FS* cases) lead to "R-NXT" as 2.41 words in successive steps, however, concerning active choices, both MRES and CBE led to fewer word repetitions, especially for MRES, resulting in "R-NXT" as 1.74 in *KW* and 1.92 in *FS*. In fact, this phenomenon fits

almost all cases [5], particularly in *D* and *T* scenarios in which referential ambiguities prominently appear. This could be linked to the fact that, unlike computational models, humans are less efficient at keeping a long-term memory of the past co-occurring records and hence "R-NXT" is more preferred by humans but not for our models.

It should be noted that in our experimental scenarios, especially in *D* and *T* cases, the same features (either shape or color) from different objects could appear in a double or a triple. We measured this phenomenon by computing *within-trial feature repetition* ("R-WHT") as the average number of repetition of the same features from samples within a trial. These repetitions have the effect of simply reducing the complexity of each trial and is extensively used by our active learning strategies. For instance in the *T* scenario, "R-WHT" is 0.88 (in *FS*) and 0.90 (in *KW*) with the random strategy and 2.02 (in *FS*) and 2.04 (in *KW*) with the MRES. This is in contrast with [20], where such "R-WHT" is impossible, every trial consisting in four mutually different objects.

It is also interesting to observe (by referring to the results in Table II and Figure 4) the tendency that the larger the "R-WHT", the better the learning performance is. As a conclusion, within-trial repetition is exploited in priority by an effective active learning strategy to reduce ambiguity.

B. Case of prevented within-trial repetition

Although a comparison of learning behaviors has been conducted between the behavior of our proposed models and humans, the within-trial repetition was strongly exploited in our experiments by active learners. Hence, in order to conduct a more effective comparison with the results in [20], we set restrictions on all active learning experiments in Section V-B such that "R-WHT" is inhibited in each trial. The learning progress using *FS* data is illustrated in Figure 7; the performance gains for active learning versus random learning measured by the area under the curve are shown in Table III, and repetition statistics are recorded in Table IV.

TABLE III
PERFORMANCE GAIN OF ACTIVE LEARNING OVER RANDOM LEARNING IN "T2IMG" WHEN "R-WHT" IS INHIBITED.

Case \ Strategy		**MRES** over **ramdom**		**CBE** over **ramdom**	
		D	***T***	***D***	***T***
R-WHT inhibited	*KW*	0.003	0.022	0.01	0.004
	FS	0.033	0.022	0.006	0.011

From these experiments, we observe in Table III a sudden fall of performance, up to 40% compared to the experiments where "R-WHT" is possible, in terms of learning speed, along with an obvious shrink of the gain difference between active and random learning. However, active choice still slightly outperforms random choice, and an increase of "R-NXT" in the process of active learning can be observed by comparing Table II and IV.

[5]despite the fact that an exception occurs when MRES is applied in *FS* with *S* data

TABLE IV
WORD REPETITIONS MEAN VALUE IN SUCCESSIVE TRIALS WHEN "R-WHT" IS INHIBITED.

Case \ Repetition / Strategy		**MRES**	**CBE**	**random**
		R-NXT	R-NXT	R-NXT
KW	***S***	0.30	0.32	0.32
	D	1.18	1.24	1.27
	T	2.74	2.83	2.80
FS	***S***	0.33	0.33	0.32
	D	1.27	1.27	1.25
	T	2.85	2.84	2.84

As for the conclusion in [20] regarding the fact that active human learners who repeat multiple pairs (of word and object) perform better than those that repeat just one pair, the data in Table IV partially agree with it. However, despite these statistics not being significantly different from random sample selection, computational learning models use comparatively less immediate repetitions than humans.

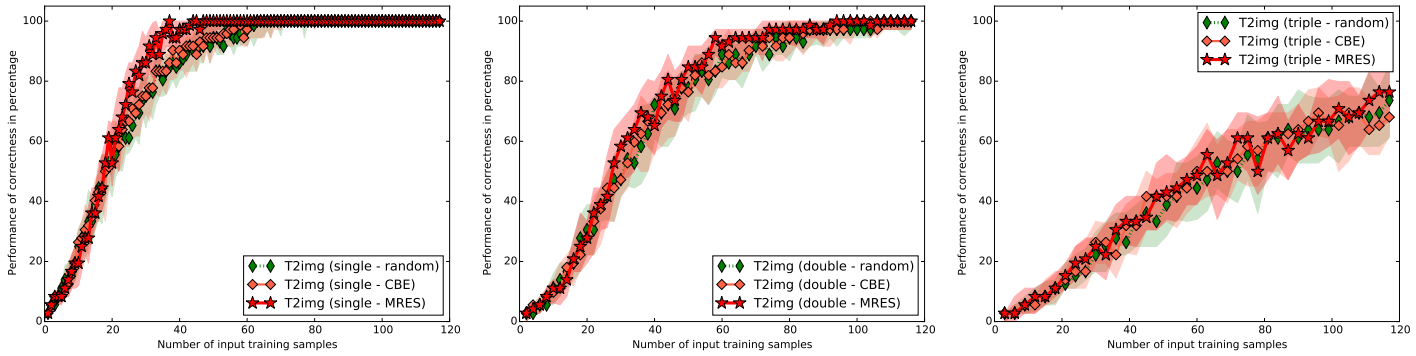
C. Discussion on the active learning behavior

Finally, we go back to the summary at behavioral level about the learning performance. First of all, "R-WHT" plays a predominant role in disambiguation, and acts as the main power to make active learning effective for computational algorithms. Note that while it is not shown in Kachergis et al., it could also be consistent with human learning experience. For example, it is shown in [52] that giving multiple samples that share similarities concerning category within a trial leads to better conceptual understanding compared to giving just a single sample to human participants. This behavior is also a more plausible scenario for a human actively asking samples with multiple objects. Indeed, a human would probably more naturally ask for object with some features in common and take advantage of feature repetition (e.g., present several red objects).

Furthermore, immediate repetition ("R-NXT"), while extensively used by humans, is used almost at the same level for both active and random learning when "R-WHT" is inhibited. This difference does not only shed little light on the relation between human and algorithmic active learning, but mainly highlights the important role of repetitions to compensate for the relatively low performance of human memory as compared to learning algorithms that can exploit the disambiguation via the repetition of features seen long ago.

VII. CONCLUSION AND FUTURE WORK

Inspired by the infants' ability to learn to recognize and name objects during parent-child interactions, two *cross-situational learning* models associated with active learning strategies were investigated: Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Association (LDA). These strategies are Maximum Reconstruction Error based Selection (MRES) and Confidence Base Exploration (CBE). As an extension of the complete computational models proposed in [17], [18], a systematic analysis of the effects of active learning was conducted.



(a) "T2img" results using $FS\&S$ (b) "T2img" results using $FS\&D$ (c) "T2img" results using $FS\&T$

Fig. 7. Comparisons in performances between MRES, CBE and random choice by applying LDA with *FS* data when "R-WHT" is inhibited.

According to the experimental results reported here, evaluated with an image recognition task that simulates the everyday testing of a learner's achievements, both active learning strategies – MRES and CBE – prove to be well-adapted to the proposed cross-situational models. Although they exhibit slight superiority over random sample selection in almost all cases in terms of learning speed and overall quality, they only show clear advantage in the more ambiguous scenarios. Analyses of the models' behavior were also performed, first by observing the *immediate repetition* of features between trials, which was observed in [20]. This led to the result that active learning algorithms show less use of *immediate repetitions* while applying more *within-trial feature repetitions* than humans. As for the preference of repeating multiple word-object pairs rather than a single pair, as concluded in [20] for humans, this paper to some extent supports it, and indicates that computational learning models use comparatively less immediate repetitions than humans, who have to rely on short-term memory.

In future work, we imagine extending our work to more complex datasets, including more diverse objects and a larger number of words. This would probably be more straightforward through the use of LDA which is more efficient in dealing with the language noise, but would require a better object representation so that the clustering required by LDA remains efficient. We plan to use features learned through deep learning [53] for this objective.

The proposed cross-situational learning models should be further compared to other word-referent learning approaches, on the one hand to models that have the goal of modeling human behaviors [54], and on the other hand to models, such as Canonical Correlation Analysis, used for the closely related task of image captioning [55].

Finally, we plan to extended our model to capture synonymous and polysemous relationships as well as to have a hierarchical extension of NMF [56] or LDA [57] to tackle the nested categories of concepts (e.g., Animals $\Rightarrow$ Mammals $\Rightarrow$ Dogs). In addition, more comparative studies concerning learning behaviors between artificial learner and real humans can be conducted to shed light on the essence of active learning mechanisms. In particular, implementing a short term memory in our learning algorithms would be important to further study the sample repetition effects.

APPENDIX
PERFORMANCES OF LEARNING USING *KW* DATA

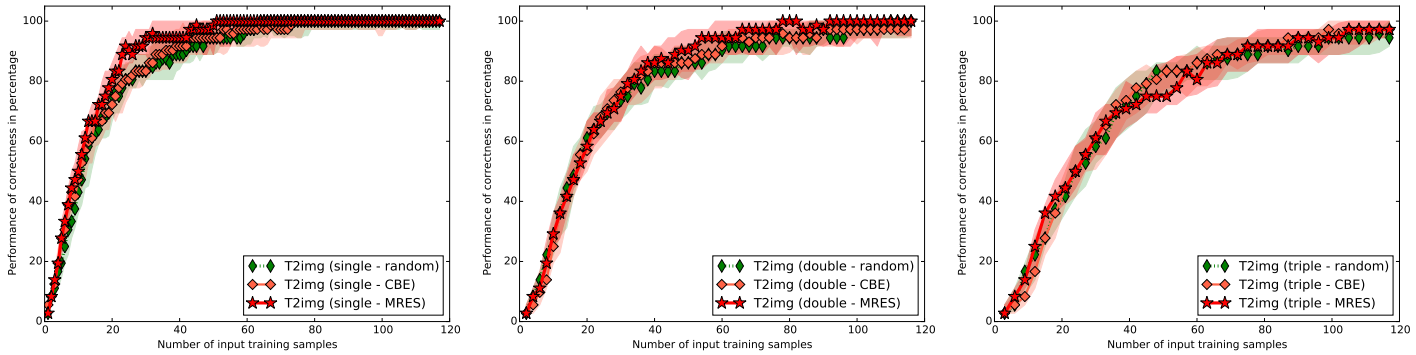
Figures 8 and 9 show the performances of NMF and LDA with samples using only keywords.

ACKNOWLEDGMENT

This work is supported by the China Scholarship Council.

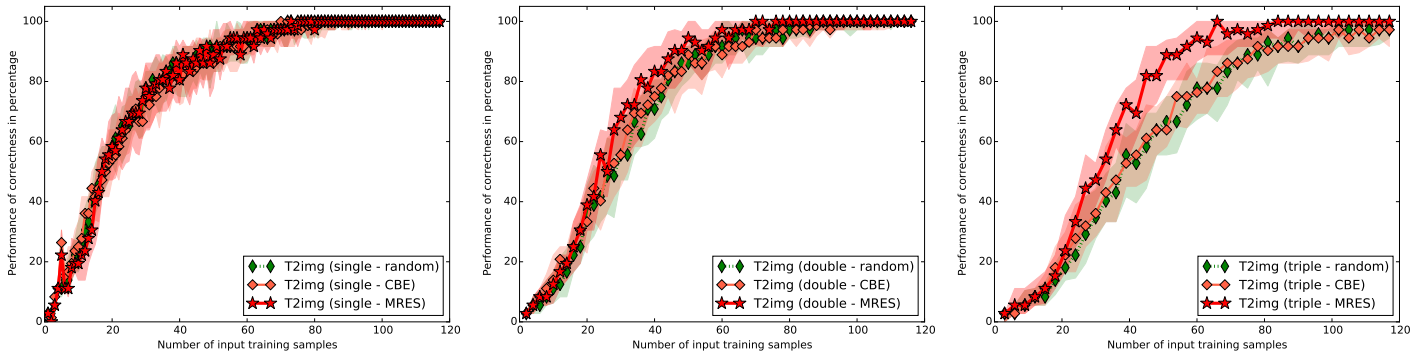
REFERENCES

[1] W. V. O. Quine, O. Neurath, and J. G. Miller, "Word and Object," Language, pp. 1–201, 1960.
[2] S. Pinker, Language Learnability and Language Development, 1984, vol. 193.
[3] N. Akhtar and L. Montague, "Early lexical acquisition: the role of cross-situational learning," pp. 347–358, 1999.
[4] A. Bandura and R. H. Walters, Social learning and personality development, 1963.
[5] A. Bandura, "Social learning theory," in Social Learning Theory, 1971, pp. 1–46.
[6] J. F. Cangelosi and A., "Cross-situational and supervised learning in the emergence of communication," Interaction Studies, vol. 12, no. 1, pp. 119—-133, 2011.
[7] J. Piaget, Play, Dreams and Imitation in Childhood, ser. Developmental psychology. Routledge, 1999. [Online]. Available: http://books.google.fr/books?id=FsdMQfpw9z0C
[8] M. Tomasello, The Cultural Origins of Human Cognition, 1999, vol. 114.
[9] J. Call and M. Carpenter, "Three sources of information in social learning," in Imitation in animals and artifacts, 2002, pp. 211–228.
[10] D. H. Grollman and O. C. Jenkins, "Sparse incremental learning for interactive robot control policy estimation," in 2008 IEEE International Conference on Robotics and Automation, 2008, pp. 3315–3320.
[11] S. Calinon and A. G. Billard, "What is the Teacher ' s Role in Robot Programming by Demonstration? Toward Benchmarks for Improved Learning," Science, vol. 8, pp. 441–464, 2007.
[12] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," Artificial Intelligence, vol. 172, pp. 716–737, 2008.
[13] C. L. Isbell, C. R. Shelton, M. Kearns, S. Singh, P. Stone, P. Avenue, and F. Park, "A Social Reinforcement Learning Agent," Fifth International Conference on Autonomous Agents, p. 8, 2001.
[14] R. Maclin, J. Shavlik, L. Torrey, T. Walker, and E. Wild, "Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression," in Proceedings of the 20th National Conference on Artificial Intelligence, vol. 2, 2005, pp. 819–824.
[15] P. Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," IEEE Transactions on Evolutionary Computation, vol. 11, no. 2, pp. 265–286, April 2007.



(a) "T2img" results using $KW\&S$ (b) "T2img" results using $KW\&D$ (c) "T2img" results using $KW\&T$

Fig. 8. Comparisons in performances between MRES, CBE and random choice by applying NMF with *KW* data.



(a) "T2img" results using $KW\&S$ (b) "T2img" results using $KW\&D$ (c) "T2img" results using $KW\&T$

Fig. 9. Comparisons in performances between MRES, CBE and random choice by applying LDA with *KW* data..

[16] P. Chandrashekhariah, G. Spina, and J. Triesch, "Let it Learn-A Curious Vision System for Autonomous Object Learning." in VISAPP (2), 2013, pp. 169–176.

[17] Y. Chen and D. Filliat, "Cross-situational noun and adjective learning in an interactive scenario," in ICDL-Epirob, Providence, United States, Aug. 2015.

[18] Y. Chen, J.-B. Bordes, and D. Filliat, "An experimental comparison between nmf and lda for active cross-situational object-word learning," in ICDL EPIROB, 2016.

[19] G. Kachergis, C. Yu, and R. M. Shiffrin, "Temporal contiguity in cross-situational statistical learning," 2009.

[20] ——, "Actively learning object names across ambiguous situations," topiCS, vol. 5, no. 1, pp. 200–213, 2013.

[21] P. J. Gorniak, "The affordance-based concept," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.

[22] S. Harnad, "The symbol grounding problem," Physica D: Nonlinear Phenomena, vol. 42, no. 1, pp. 335–346, 1990.

[23] L. Steels, "Evolving grounded communication for robots," Trends in Cognitive Sciences, vol. 7, no. 7, pp. 308 – 312, 2003.

[24] S. D. Larson, Intrinsic representation: Bootstrapping symbols from experience. Springer, 2004.

[25] C. Yu, L. B. Smith, and A. F. Pereira, "Grounding word learning in multimodal sensorimotor interaction," in Proceedings of the 30th annual conference of the cognitive science society, 2008, pp. 1017–1022.

[26] N. Mavridis, C. Datta, S. Emami, C. BenAbdelkader, A. Tanoto, and T. Rabie, "Facebots: social robots utilizing facebook," in Proceedings of the 4th ACM/IEEE international conference on Human robot interaction. ACM, 2009, pp. 195–196.

[27] K. Gold and B. Scassellati, "Grounded pronoun learning and pronoun reversal," in Proceedings of the 5th International Conference on Development and Learning, 2006.

[28] D. K. Roy, "A computational model of word learning from multimodal sensory input," in Proceedings of the International Conference of Cognitive Modeling (ICCM2000), Groningen, Netherlands. Citeseer, 2000.

[29] N. Mavridis and D. K. Roy, "Grounded situation models for robots: Where words and percepts meet," in Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. IEEE, 2006, pp. 4690–4697.

[30] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Krose, "Biron, where are you? enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization," in Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. IEEE, 2006, pp. 934–940.

[31] T. Regier and L. A. Carlson, "Grounding spatial language in perception: an empirical and computational investigation." Journal of experimental psychology: General, vol. 130, no. 2, p. 273, 2001.

[32] D. K. Roy, "Learning visually grounded words and syntax for a scene description task," Computer Speech & Language, vol. 16, no. 3, pp. 353–385, 2002.

[33] K. R. Coventry and S. C. Garrod, Saying, seeing and acting: The psychological semantics of spatial prepositions. Psychology Press, 2004.

[34] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 34, no. 2, pp. 154–167, 2004.

[35] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," Robotics and Autonomous Systems, vol. 56, no. 6, pp. 493–502, 2008.

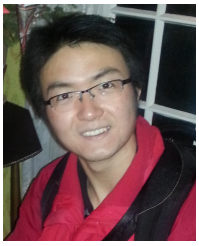
[36] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Approaching the symbol grounding problem with probabilistic graphical models," AI magazine, vol. 32, no. 4, pp. 64–76, 2011.

[37] S. Tellex, P. Thaker, J. Joseph, and N. Roy, "Learning perceptually grounded word meanings from unaligned parallel data," Machine Learning, vol. 94, no. 2, pp. 151–167, 2014.

[38] S. Coradeschi, A. Loutfi, and B. Wrede, "A short review of symbol grounding in robotic and intelligent systems," KI-Künstliche Intelligenz, vol. 27, no. 2, pp. 129–136, 2013.

[39] P. Vogt, "The physical symbol grounding problem," Cognitive Systems Research, vol. 3, no. 3, pp. 429–457, 2002.

[40] A. Cangelosi, "The grounding and sharing of symbols," Pragmatics & Cognition, vol. 14, no. 2, pp. 275–285, 2006.
[41] C. Yu and D. H. Ballard, "On the integration of grounding language and learning objects," in AAAI, vol. 4, 2004, pp. 488–493.
[42] O. Mangin, "The Emergence of Multimodal Concepts," Ph.D. dissertation, 2014.
[43] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, H. van den Heuvel et al., "A speech corpus for modeling language acquisition: Caregiver." in LREC, 2010.
[44] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi, "Autonomous acquisition of multimodal information for online object concept formation by robots," in IEEE International Conference on Intelligent Robots and Systems, 2011, pp. 1540–1547.
[45] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," Robotics and Autonomous Systems, vol. 62, no. 6, pp. 721–736, Jun. 2014.
[46] D. K. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," Cognitive science, vol. 26, pp. 113–146, 2002.
[47] T. M. Cover and J. A. Thomas, Elements of information theory. John Wiley & Sons, 2012.
[48] W. Schueller and P.-Y. Oudeyer, "Active learning strategies and active control of complexity growth in naming games," in the 5th International Conference on Development and Learning and on Epigenetic Robotics, 2015.
[49] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Aug. 1988.
[50] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." Nature, vol. 401, pp. 788–791, 1999.
[51] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937
[52] D. Gentner and L. L. Namy, "Comparison in the development of categories," Cognitive development, vol. 14, no. 4, pp. 487–513, 1999.
[53] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 53–61.
[54] G. Kachergis, C. Yu, and R. M. Shiffrin, "A bootstrapping model of frequency and context effects in word learning," Cognitive Science, vol. 41, no. 3, pp. 590–622, 2017. [Online]. Available: http://dx.doi.org/10.1111/cogs.12353
[55] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," Int. J. Comput. Vision, vol. 106, no. 2, pp. 210–233, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1007/s11263-013-0658-4
[56] K. Kersting, M. Wahabzada, C. Thurau, and C. Bauckhage, "Hierarchical Convex NMF for Clustering Massive Data." ACML, pp. 253–268, 2010. [Online]. Available: http://jmlr.org/proceedings/papers/v13/kersting10a/kersting10a.pdf
[57] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in Advances in Neural Information Processing Systems. MIT Press, 2004, p. 2003.



Yuxin Chen received the master's degree in the School of Electronic Information and Electrical Engineering (SEIEE) from Shanghai Jiao Tong University, China in 2013 and Ph.D degree in the graduate school of Approches Interdisciplinaires, Fondements, Applications et Innovation (INTERFACES) from l'Université Paris-Saclay in 2017. The Ph.D research project is under the joint-supervision of Unité d'informatique et d'ingénierie système (U2IS) from L'ÉNSTA-ParisTech and the FLOWing Epigenetic Robots and Systems (Flowers) at Inria. His research interests focus on interactive learning mechanisms for developmental robotics, in the interdisciplinary area of computer vision, machine learning, multimodal information processing and active learning strategies.



Jean-Baptiste Bordes graduated from Ecole Polytechnique in 2004 and obtained a PhD in computer vision from Telecom ParisTech in 2009. After 3 years as a data architect for information systems in the French armament procurement agency, he worked as a research engineer at the Université de Technology de Compiègne (UTC) on the topic of perception for autonomous vehicles until 2011. Until now, he is Dean of Studies for the Engineer Program of Ecole Polytechnique. His main research interest are perception and learning in the frame of developmental approach for autonomous mobile robotics.



David Filliat graduated from the Ecole Polytechnique in 1997 and obtained a PhD on bio-inspired robotics navigation from Paris VI university in 2001. After 4 years as an expert for the robotic programs in the French armament procurement agency, he is now professor at Ecole Nationale Supérieure de Techniques Avancées ParisTech. Head of the Robotics and Computer Vision team since 2006, he obtained the Habilitation à Diriger des Recherches en 2011. He is also a member of the ENSTA ParisTech INRIA FLOWERS team. His main research interest are perception, navigation and learning in the frame of the developmental approach for autonomous mobile robotics. http://www.ensta-paristech.fr/~filliat/