Cogn Process
DOI 10.1007/s10339-011-0407-y

RESEARCH REPORT

Using spoken words to guide open-ended category formation

Aneesh Chauhan · Luís Seabra Lopes

Received: 17 September 2010 / Accepted: 3 May 2011

Abstract Naming is a powerful cognitive tool that facilitates categorization by forming an association between words and their referents. There is evidence in child development literature that strong links exist between early word-learning and conceptual development. A growing view is also emerging that language is a cultural product created and acquired through social interactions. Inspired by these studies, this paper presents a novel learning architecture for category formation and vocabulary acquisition in robots through active interaction with humans. This architecture is open-ended and is capable of acquiring new categories and category names incrementally. The process can be compared to language grounding in children at single-word stage. The robot is embodied with visual and auditory sensors for world perception. A human instructor uses speech to teach the robot the names of the objects present in a visually shared environment. The robot uses its perceptual input to ground these spoken words and dynamically form/organize category descriptions in order to achieve better categorization. To evaluate the learning system at word-learning and category formation tasks, two experiments were conducted using a simple language game involving naming and corrective feedback actions from the human user. The obtained results are presented and discussed in detail.

Keywords Vocabulary acquisition · Open-ended category learning · Social language grounding · Human-Robot interaction · Embodied agents Conceptual development · Instance-based learning · Word categories · Object categories

Introduction

Words are at the core of the language understanding and acquisition processes. Before infants are able to grasp the rules associated with their native languages, they are already accumulating lexicon at a gradually increasing rate. Vocabulary in children starts with about 10 words in the first year and reaches to almost 300 words by the end of the second year (Bates et al. 1992; Bloom 2000; Crystal 1987; Fenson et al. 1994). However, the process of vocabulary acquisition (and thus language acquisition) has its foundation at an even earlier age. There is mounting evidence that the process of language acquisition begins when infants start learning the sounds of words (language specific phonemes) and are able to distinguish these sounds from the auditory stream (Jusczyk 1993; Yeung and Werker 2009). By the age of 9–12 months, infants are able to individuate familiar words from speech and begin building a lexicon (Jusczyk and Aslin 1995; Waxman 2008). This paper views vocabulary acquisition in robots from the perspective of these studies.

As a baseline, this paper only assumes that the robot is already at a stage where it can detect the phonemes in a spoken word (6 months of age in human time scale). This stands in contrast to more conventional approaches that do

The Portuguese Foundation for Science and Technology (FCT) supported this work under contract: SFRH/BD/31577/2006 (PhD scholarship grant to the first author).

A. Chauhan (✉) · L. Seabra Lopes
Instituto de Engenharia Electrónica e Telemática
de Aveiro (IEETA), Universidade de Aveiro,
Aveiro 3810-193, Portugal
e-mail: aneesh.chauhan@ua.pt

L. Seabra Lopes
e-mail: lsl@ua.pt

Published online: 26 May 2011

not take the sounds of words into account. These approaches either use words directly transmitted in text (e.g., Cangelosi and Harnad 2000; Chauhan and Seabra Lopes 2010; Seabra Lopes and Chauhan 2008) or process spoken words using speech recognition tools but ignore the respective recognition uncertainty (e.g., Gold et al. 2009; Krunic et al. 2009; Levinson et al. 2005; Skočaj et al. 2007) There are, however, some notable exceptions. Yu and Ballard (2003, 2004, 2007) treat a phoneme sequence as a string and use a string matching algorithm (using string changing operations, such as insertion and deletion) to measure the amount of difference between two phoneme sequences. Roy and Pentland (2002, 2003) generate a Hidden–Markov Model (HMM) from the phoneme sequence predicted for a given speech segment, where each phoneme is assigned an HMM state. The HMM state transitions are strictly left-to-right and the transition probabilities are given by the phoneme models previously trained on a context-independent dataset. To compare two speech segments, they proposed a distance metric which computes the likelihood of producing one speech segment given the HMM of the other speech segment.

In this paper, the approach to word representation is based on two sets of features extracted from the word utterance: the sequence of predicted phonemes; and the Mel-frequency cepstrum (MFC) set. The raw speech signal (spoken word) is represented by the sequence of predicted phonemes, where each phoneme is coupled to a set of Mel-frequency cepstral coefficients (MFCC) computed over the same time-frame for which the phoneme was predicted. To measure the similarity between two word representations, we use a greedy search algorithm to find a locally optimal alignment between them. The feature extraction process, the word representation, and the similarity measures are detailed in section "Learning and classification".

In itself, a word does not contain a meaning. Its meaning lies in its association with the entity of the world it refers to Barsalou (1999), Harnad (1990). The problem of associating a word (or a symbol) to its referent is known as the symbol grounding problem (Harnad 1990). Robotic simulations of language origins, evolution, and transfer have provided powerful evidence that language is grounded socially (Roy and Pentland 2002; Loreto and Steels 2007; Steels 2008; Steels and Kaplan 2002). These studies bring new insights into language emergence and acquisition in humans. A new view among linguists is also gaining strength which considers language a cultural product which is transmitted and grounded via social interactions (Cowley 2007; Love 2004). Studies on child language development have also pointed out that naming has a powerful influence on conceptual organization and categorization (Waxman 2008; Yoshida and Smith 2005). Naming reinforces children's learning of the links between perceptual cues (e.g., shape) and category formation (Yoshida and Smith 2005).

Inspired by these studies, an approach to grounding vocabulary through social transfer is presented, where a human, acting as an instructor, teaches a robotic agent the names of the objects present in their visually shared environment. It is deemed essential that, if robots and humans are to share a common language, humans must act as instructors/mediators (Seabra Lopes and Chauhan 2008; Steels and Kaplan 2002; Seabra Lopes and Chauhan 2007; Thomaz and Breazeal 2008; Thomaz et al. 2006).

In standard approaches to grounding vocabulary in artificial agents, object category formation is taken as a supervised learning task that assists in vocabulary grounding (Chauhan and Seabra Lopes 2010; Seabra Lopes and Chauhan 2008; Gold et al. 2009; Krunic et al. 2009; Skočaj et al. 2007; Steels and Kaplan 2002; Seabra Lopes and Chauhan 2007). This approach makes sense when words are communicated reliably, e.g., as text. However, when words are communicated via speech, with all constraints usually associated to speech communication (speaker voice, accent, environment noise, etc.), different utterances of the same word display some amount of variation, leading to possible confusion between utterances of different words. Thus, on the listener's side, interpreting a spoken word involves, firstly, to recognize the word itself, i.e., to map the word to a previously known category of spoken words (a word category). The next interpretation step is semantic interpretation, or grounding, understood as the association of a meaning to the heard word. Roy and Pentland (2002, 2003) developed a system (CELL) that discovers relevant linguistic units in speech input as well as relevant visual categories. Co-occurring visual categories and linguistic units are paired together to form the so-called AV events (audio-visual events). AV events are consolidated in memory through clustering leading to lexical items. Yu and Ballard (2003, 2004, 2007) have developed a system that also identifies "word-like units" in an utterance and co-occurring meanings in a visual scene. Pairing of words and meanings is achieved through the Expectation-Maximization algorithm.

This paper presents a novel learning and categorization approach to spoken word grounding for artificial cognitive agents. The learning approach is open-ended in that there is no set of words and meanings defined in advance, and new words and meanings are acquired incrementally through interaction with a human instructor. The open-ended nature of the approach was already present in our previous work (Chauhan and Seabra Lopes 2010; Seabra Lopes and Chauhan 2007, 2008). However, that previous work focused on grounding textually communicated words. The approach was now extended to support spoken word grounding, as described in this paper. In this new work, we assume that most word categories are reasonably homogeneous when compared to object categories (meanings). This has led us to explore mechanisms that use word categories to dynamically form and reorganize object categories, while word categories themselves are formed and reorganized through clustering of word instances.

The early lexicon in children (till the age of 3) consists mainly of common nouns that name concrete objects in the children's environment (e.g., toys, food items, geometric shapes, animal categories) and to a lesser extent routine social words, proper nouns, animal sounds, and observable verbs (Bloom 2001; Messer 1994). The overwhelming bias toward learning names of concrete object categories is a direct consequence of the early conceptual development process in infants. Studies have shown that infants at a very young age start showing an attentional bias toward categories that have clearly defined shapes (Bomba and Siqueland 1983; Landau et al. 1988; Smith and Samuelson 2006). This makes grounding the names of visually concrete referents an easier task and their early existence prevalent in comparison to other words (Gillette et al. 1999).

The words taught to the robotic agent presented in this paper also refer to concrete objects normally present in children's environments. Other computational models for studying the word-learning process have also followed similar inspiration. For example, several researchers teach the names of children's toys (Levinson et al. 2005; Roy and Pentland 2002; Steels and Kaplan 2002), others teach the names of office objects (Seabra Lopes and Chauhan 2007) or toys, cutlery, and office objects (Seabra Lopes and Chauhan 2008).

Our approach is implemented and tested on an embodied agent with physical components to help interact with its users and its environment, namely a camera, a microphone, and a robotic arm. Its action capabilities range from linguistic and visual responses to the manipulation of objects by the robotic arm. The types of features used to describe objects, the visual feature extraction process and the methods for classification of object categories are taken from our previous works (Chauhan and Seabra Lopes 2010; Seabra Lopes and Chauhan 2008; Seabra Lopes et al. 2007).

Rest of the paper is organized as follows: Next section describes the paradigm for human-robot interaction, which was designed to facilitate language transfer from a human user to the robot. Two sections that follow, detail the signal representation and representation strategy, and the learning and categorization model of the robot, consecutively. A set of experiments based on a simple teaching protocol is reported and the results are discussed in "Experimental evalution". The final section summarizes the approach and concludes the paper.

2 Human–robot communication interface

As a cognitive tool, the primary purpose of a language is to communicate about the entities of the world. Meaning formation, on the other hand, is a cognitive task of representing these entities in an individual's brain. Although linked, language (as a communication tool) and the formation of meaning are two separate cognitive tasks. This distinction is extremely important because it demonstrates that any two individuals share a language when they have the same words grounded to the same entities, regardless of their respective processes of meaning formation. This is exploited here for teaching a human language to a robotic agent through human assistance. Meaning formation for a robot cannot be directly compared with that for humans (Steels 2008). Still, robots can learn a human language if they can ground human language symbols (words) in their perception of the world.

In this work, a human instructor uses speech to teach the words (language transfer) in her/his own language (English in this case) to a robot that has its own perception mechanism. The human instructor teaches the robot the names of the objects present in a visually shared dynamic environment. This environment is basically a black table top, where the black color was chosen to simplify object detection and extraction. The scene is captured through a camera placed over the table. Object names are then grounded by the robot in sensor-based object descriptions, leading to a vocabulary shared with its instructor. Shared attention is deemed essential if words are to be correctly grounded. In the described work, shared attention is achieved when the human user, by mouse-clicking, selects an object from the robot's visual scene (camera frame).

Once shared attention is established, the instructor can interact with the robot by performing the following actions (selected from a menu):

1. *Teach*: teach the category name of the selected object (simultaneous visual and vocal input);
2. *Ask*: ask the category name of the selected object, which the agent will predict based on previously learned knowledge (visual input); and
3. *Correct*: if the category predicted in the previous case is incorrect, the user can teach the correct category (vocal input).

By recording and reproducing the audio information given through *teach* and *correct* actions (recordings of

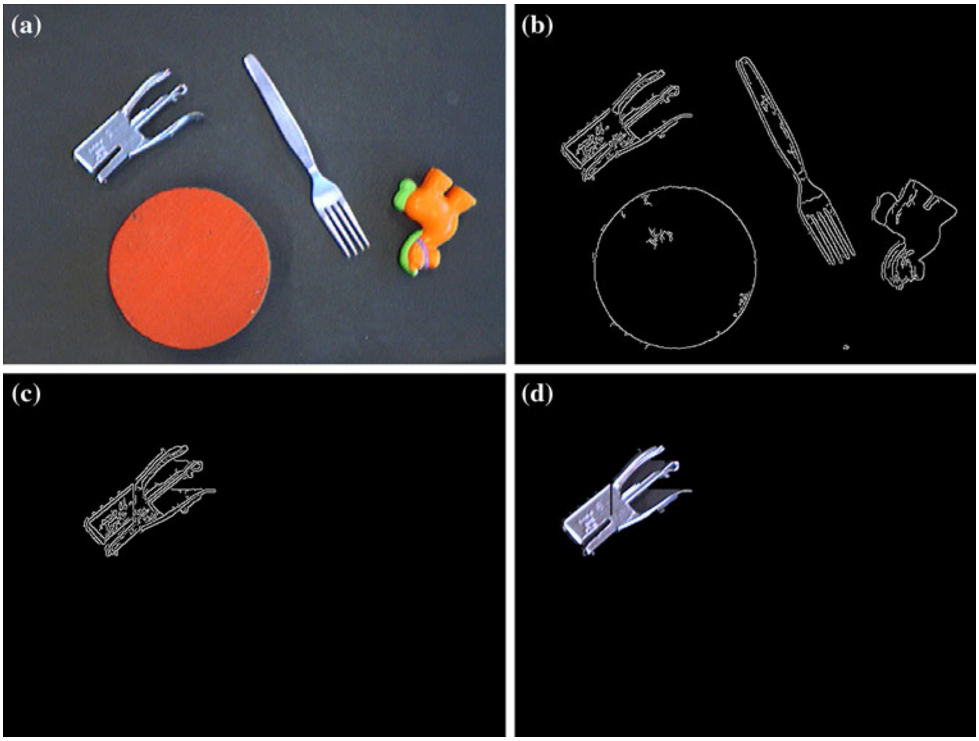

Fig. 1 Sequence of steps involved in the extraction of object "*Stapler*", as selected by the user, from the robot's visual scene. **a** Original scene with four objects; **b** Edges-based counterpart of the visual scene; **c** Edges of the selected object ("*Stapler*"); **d** "*Stapler*" extracted from the original scene

words), the robot is able to respond to *ask* actions from the human interlocutor.

3 Signal perception and representation

The robotic agent is embodied with a cheap off-the-shelf microphone and an IEEE1394 compliant firewire camera, respectively for auditory and visual perceptions. The raw signals are processed to extract auditory and visual representations which will then be used for learning and classification.

3.1 Scene images

When the user points the mouse to an object in the current scene image, an edge-based counterpart of the whole image is generated. From this edges image, the edges of the object are extracted, taking into account the user-pointed position, and assuming that different objects do not occlude each other in the image. Finally, given the edges of the object, an image of the object is extracted from the original scene image (segmented object) (see Fig. 1). From the segmented object and its edges, different representations are extracted.

The visual feature extraction and representation module has been borrowed directly from our previous work (Seabra Lopes and Chauhan 2008). An object is represented in multiple feature spaces. Most of them are designed to capture shape information. Different feature spaces capture different aspects of a particular object and are possibly complementary to each other. In total, 7 feature spaces are used:

- Five feature spaces extracted from the edge-based counterpart of the object image. These features spaces are obtained by segmenting the smallest circle enclosing the edges image of the object and centered in its geometric center. For different feature spaces, such a circle is segmented either into a number of slices or a number of concentric layers. These feature spaces are translation, rotation, and scale invariant.
- One feature space is composed of a single feature, the area, defined as the total number of pixels of the object in the image. This is the only scale-dependent feature.
- One feature space is used to describe objects in terms of their main colors (HSV color space is used) (Seabra Lopes et al. 2007).

3.2 Vocal sounds

Vocal signals are collected in time slots of fixed duration. One of the reasons for setting a limit on the length of the spoken word is that the software (SoX[1]) used for capturing the audio signal from the microphone, requires either an explicit input from the user (Ctl+C signal from the keyboard) to stop recording, or a time-frame can be defined

[1] http://sox.sourceforge.net/.

after which SoX halts the recording. We chose the second method to maintain the autonomy of the agent. After crude testing, the time-frame for word utterances has been set to two seconds in the current implementation. This limit allows a user, unfamiliar with the agent, to conveniently speak a single word within the time limit. A serious drawback of this approach is that, for the same word, different recordings can be out of synchronization. But our hypothesis is that the measure used to compute the similarity between two word representations (see "Learning and classification") will be invariant to small unsynchronizations.

From the raw audio signal, two sets of features are extracted: the phoneme sequence; and the Mel-frequency cepstrum (MFC) set. MFC provides a good approximation to the response of the human auditory system to a sound stream. This set is obtained using the *wave2feat* tool provided with Sphinx3[2] speech recognition engine. Using this tool with the standard settings, the 2s speech signal is divided into 199 equal sized speech segments. For each of these segments, MFC is calculated, and the 10 most significant initial amplitude values (*i.e.*,. the first 10 Mel-frequency cepstral coefficients, or MFCC) are stored.

To extract the phoneme sequence, *allphone* mode of Sphinx3[3] is used. This mode predicts:

- the most probable sequence of phonemes for a given speech signal;
- the elements of the MFC set associated with each predicted phoneme.

Once the sound features have been extracted, a word W is represented as:

$$W = \{ <ph_1, m_1> , <ph_2, m_2> , .., <ph_n, m_n> \}$$

where n is the number of predicted phonemes, ph_i is the i-th phoneme in the sequence, and m_i is the set of all MFCC vectors in the time period for which ph_i was predicted (thus, m_i is a subset of the 199 MFCC vectors initially computed).

To maintain speaker independence, the Sphinx3 speech recognition engine was not trained on any particular individual or for any specific vocabulary. However, this has a drawback that the predicted phoneme sequence for any spoken word contains a fairly high amount of noise. Therefore, the word representation approach described above uses both the predicted phonemes and the associated MFCC vectors. The underlying assumption is that, in case the phonemes predicted are unreliable, the corresponding MFCC vectors will provide complementary information which can compensate for the incorrect phoneme prediction.

3.3 Category representations

Instance-based representations are adopted both for spoken word categories and visual object categories. That is, each spoken word category is described by a set of representations of known word instances. Likewise, object categories are described by sets of representations of known instances. Each word category and the corresponding object category are coupled together such that each instance in the object category is associated to a word instance in the respective word category. For clarity of understanding, Fig. 2 summarizes the signal perception and representation scheme.

4 Learning and classification

This section introduces a novel strategy, which has been designed to facilitate object category formation by taking the association between names and their meanings into account. This strategy uses the information contained in the names (word categories) for dynamic formation of meanings (object categories). Before discussing the meaning formation process, it is important to present the word similarity measure as well as the word classification process.

4.1 Word similarity and classification

For the purposes of classification, a novel similarity measure has been developed for comparing any two word representations. Given two words,

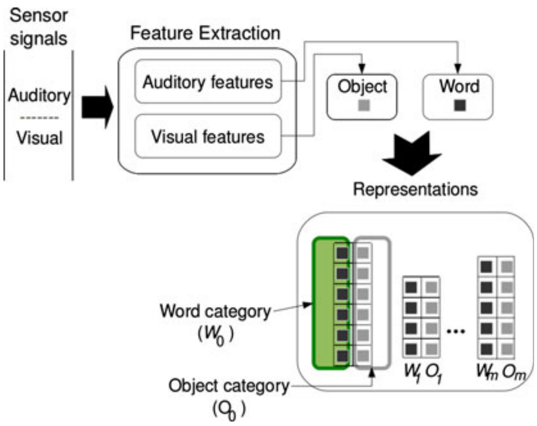

Fig. 2 Signal perception and representation schema

[2] A toolkit for speech recognition based on Hidden–Markov Models (HMM): http://cmusphinx.sourceforge.net.

[3] Trained on VoxForge, an open-source speech corpus and acoustic model repository: http://www.voxforge.org.

$$W_p = \{ <ph_{p,1}, m_{p,1}>, <ph_{p,2}, m_{p,2}>, .., <ph_{p,r}, m_{p,r}> \}$$
$$W_q = \{ <ph_{q,1}, m_{q,1}>, <ph_{q,2}, m_{q,2}>, .., <ph_{q,s}, m_{q,s}> \}$$

an algorithm based on dynamic time warping [DTW (Rath and Manmatha 2003)] is used to find the distances between each $m_{p,i}$ and $m_{q,j}$, where $i = 1, 2, \ldots, r, j = 1, 2, \ldots, s$, and r and s are the number of phonemes in the phoneme sequence of W_p and W_q, respectively. The end product of this algorithm is an $r \times s$ matrix where each element (i, j) of the matrix is the DTW distance measure for m_{pi} and m_{qj}. We will refer to this matrix as $DTW(W_p, W_q)$.

In an ideal condition, for two words in the same category, the sequences of phonemes (and respective MFCC sets) should be exactly the same. This is exemplified in the case of $DTW(W, W)$, where each diagonal element will be zero. The diagonal path provides an approximation of the match between the sequences of phonemes from two different words. The summation of diagonal values of the *DTW* matrix is one possible distance measure for comparing words. However, in real time applications, such measure will lead to a very weak performance. In alternative, a local greedy search algorithm has been designed to find a locally optimal path in the *DTW* matrix such that the sum of all the elements lead to locally optimal minimum. The objective is to reduce the total cost while maintaining proximity to the diagonal path.

Given two words W_p and W_q, this cost is given by:

$$C(W_p, W_q) = \sum_{i=1}^{min(r,s)} c_i(W_p, W_q) \qquad (1)$$

where

$$c_i(W_p, W_q) = \begin{cases} 0, if & ph_{p,i} \in \{ph_{q,i-1}, ph_{q,i}, ph_{q,i+1}\} \\ min & (DTW_{i,i-1}(W_p, W_q), \\ & DTW_{i,i}(W_p, W_q), \quad DTW_{i,i+1}(W_p, W_q)), \quad otherwise \end{cases}$$

is the cost function, which returns the distance between a diagonal element in W_p and its closest neighbor in a local search window around the corresponding element in W_q. In the special case of an exact phoneme match, it returns zero. This way, $C(W_p, W_q)$ will be zero in several relevant special cases, namely when the sequences of phonemes are identical. That wouldn't be the case if c_i would be computed based on *DTW* distances only. It is essential to note that the cost function $C(W_p, W_q)$ is not symmetric and thus not a true distance measure. Thus, the final similarity measure between two words is calculated based on the minimum cost as follows:

$$S\_word(W_p, W_q) = \frac{1}{min(C(W_p, W_q), C(W_q, W_p))} \qquad (2)$$

A classifier based on the nearest-neighbor (NN) approach has been developed for classifying words. Given an input "word" to be classified, it is compared with all the word representations stored in memory. The word category containing the instance most similar to the input word is predicted to be the category of that word.

4.2 Dynamic category formation and organization

Although the objects belonging to a certain category can be very different from each other and different object categories can share a single name, different instances in a word category will often be very similar to each other. In this work, we assume that most word categories are reasonably homogeneous and therefore should not contain more than one easily recognizable cluster of instances. Given this assumption, the presence of two or more clusters suggests that the word category actually contains instances of two or more words. Based on this, a novel methodology has been designed that uses word clustering to dynamically form/organize not only word categories, but also object categories.

The cluster identification approach involves locating the nearest-neighbor (using the *S_word* metric) of each instance, and the word category is represented as a directed graph, in which edges connect instances to their respective nearest-neighbors. One or more weakly connected components (i.e., maximal connected subgraphs computed while ignoring edge directions) will be identified in the graph. The instances in each of these components will form a different cluster (see Fig 3). For each cluster, the instance with more edges pointing to it in the graph is defined as its "key word", or w_{key}.

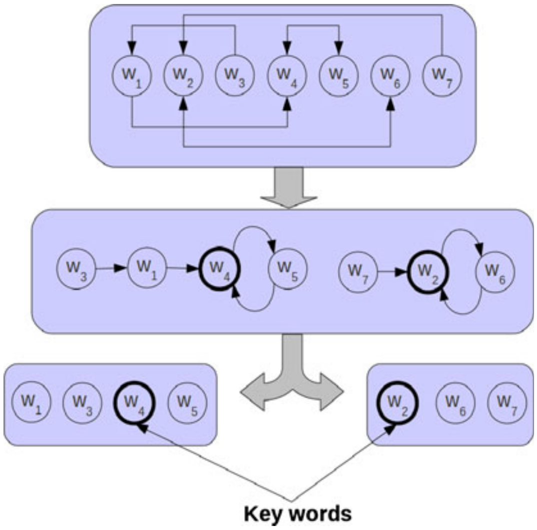

Fig. 3 Extracting clusters from a set of 7 words representing a category name (the nearest-neighbor of an instance is pointed by the head of the arrow originating from that instance); key words are the representations with the highest number of nearest-neighbor links in a given cluster

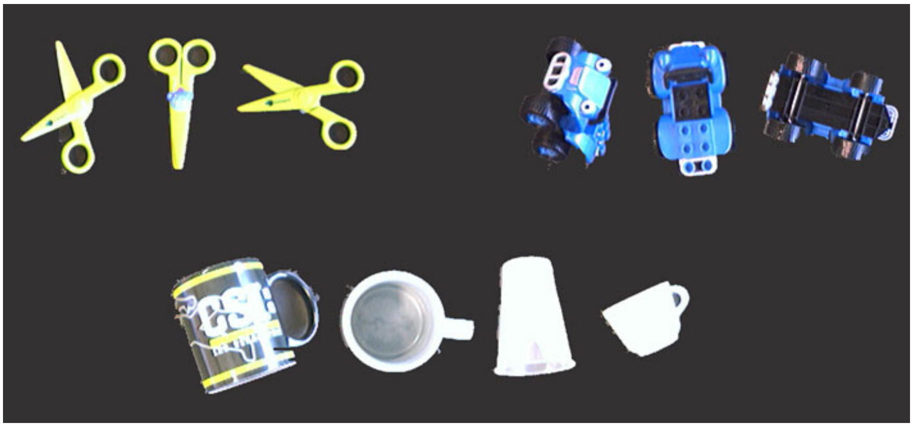

Fig. 4 A set of examples illustrating different views in which objects were shown to the robot: the instance of *scissor* in varying degrees of blade openings; the instance of *jeep* in different 3D orientations (front toward the camera, top view and bottom facing the camera); and different instances from the *cup* category in varying orientations

In the case where two or more clusters are identified for a given word category, W_i, each of them is checked if it should form a completely new word category, or if it should be merged to any of the other existing word categories. A given cluster in W_i will form a completely new word category if all instances from all other word categories are less similar to the keyword of that cluster, W_{key}, than the nearest-neighbor of W_{key} in that cluster. Otherwise, the cluster will be merged to the category W_m, $m \neq i$, containing the closest instance to W_{key}. At the same time, the associated object descriptions are moved to the corresponding object category. The complete procedure is described in Algorithm 1.

This procedure is called each time there is a *teach* or a *correct* action performed by the human user. In these situations, the user provides the category name of a selected object. This word instance (the name) is added to the word category (W_i) as predicted using the classification method described in "Word similarity and classification" and the object instance is added to the object category (O_i) coupled with that word category.

4.3 Object classification

As mentioned earlier, an instance-based approach is used for object category representation, where categories are represented by sets of known instances. New instances are stored only when there is an explicit teaching action or a corrective feedback.

The strategy for classification, adopted from our previous works (Chauhan and Seabra Lopes 2010; Seabra Lopes and Chauhan 2008; Seabra Lopes et al. 2007), is based on classifier combinations. Thus, two types of classifiers are included:

- Base classifiers: 6 nearest-neighbor (NN) classifiers (Chauhan and Seabra Lopes 2010) and 10 nearest-cluster (NC) classifiers (Chauhan and Seabra Lopes 2010), mostly based on the shape feature spaces mentioned above in "Scene images". These classifiers are based either on Euclidean similarity or on the Pyramid Match Score (Grauman and Darrell 2007). A color-based classifier is also included (Seabra Lopes et al. 2007).
- 7 classifier combinations, based on majority voting and Dempster–Shafer evidence theory (Seabra Lopes and Chauhan 2008).

A metacognitive component (Seabra Lopes and Chauhan 2008) maintains updated success statistics for all the classifiers and, based on these statistics, reconfigures classifier combinations. The final category prediction result for a given object is taken from the current most successful classifier.

The base classifiers are designed such that the categorization of a previously unseen instance involves ranking the known categories according to the measures of membership of that instance to each of the categories.

All membership measures used in base classifiers are normalizations of a similarity measure, according to the following generic formula:

$$M(y, O_i) = \frac{S(y, O_i)}{\sum_{k=1}^{N} S(y, O_k)} \qquad (3)$$

where O_i is the i-th object category, $i = 1, \ldots, N$, N is the number of categories, $S(y, O)$ is a measure of the similarity between an object y and a given object category, O. The membership values $M(y, O_i)$ sum to 1.0, allowing their use as evidence in Dempster–Shafer classifier combinations. In the color-based classifier, $S(y, O)$ is specific of that classifier and the details are omitted here (see (Seabra Lopes et al. 2007)).

In the case of NN classifiers, given an object to be classified, it is compared with all the instances stored in memory. The category containing the instance most similar to the input object is predicted as its category. Thus, in this

Algorithm 1 Dynamic category formation and organization (W, O, i)

```
$W$ - array of all word categories (input/output)
$O$ - array of all object categories (input/output)
$i$ - index of a particular category in W and O (input)

$Clusters \leftarrow$ clusters formed for $W_i$
$p \leftarrow$ number of clusters in $W_i$
if $p = 1$ then return

repeat
  $w_{key} \leftarrow$ find the key word for $Clusters[p]$
  $s \leftarrow$ max_similarity($w_{key}$, $Clusters[p]$)      ▷ the measure of similarity between $w_{key}$ and its closest neighbor
  $c \leftarrow$ number of known categories
  $m \leftarrow i$
  repeat
    if $c \neq i$ then
      $s_{max} \leftarrow$ max_similarity($w_{key}$, $W_c$)
      if $s_{max} > s$ then
        $s \leftarrow s_{max}$
        $m \leftarrow c$
      end if
    end if
    $c \leftarrow c - 1$
  until c = 0
  if $m \neq i$ then
    Add the word instances in $Clusters[p]$ to $W_m$ and the object instances coupled to these words to $O_m$
  else
    Create a new word category for the word instances in $Clusters[p]$ and form a new object category description from the object instances coupled with these words
  end if
  $p \leftarrow p - 1$
until $p = 0$
Remove $W_i$ and $O_i$
```

case, $S(y, O)$ is the maximum similarity between the object and the instances in the category:

$$S(y, O) = \max_{x \in O} S(y, x) \quad (4)$$

In the case of NC classifiers, categories are organized into clusters. Clustering involves locating the nearest-neighbor of each instance in an object category and bringing together the instances that are connected to each other through their nearest-neighbors (the method is similar to the one used for clustering words, as explained in "Dynamic category formation and organization"). When classifying a new object, average similarity measures are computed between the object and the instances in each cluster. For each category, the cluster with the highest average similarity to the target object will provide the membership score of the category. Thus, in this case, the similarity $S(y, O)$ is the maximum average similarity between y and the objects in the different clusters, $u \in O$:

$$S(y, O) = \max_{u \in O} (\text{average}_{x \in u} S(y, x)) \quad (5)$$

For the NN and NC classifiers based on the Pyramid Match Score (Seabra Lopes and Chauhan 2008, Grauman and Darrell 2007), the similarity between two objects x and y, $S(x, y)$, is directly given by the Pyramid Match Score. In the other classifiers, based on Euclidean Distance, $S(x, y)$ is given by the inverse of Euclidean Distance.

5 Experimental evaluation

5.1 Global evaluation

The performance of the learning model was evaluated on vocabulary acquisition. The human instructor follows the "teaching protocol" proposed in Seabra Lopes and Chauhan (2007, 2008). As dictated by the protocol, the user

Algorithm 2 Teaching protocol for performance evaluation

```
Introduce $Category_1$;
$n \leftarrow 1$
repeat
    $n \leftarrow n + 1$                    ▷ Ready for the next category
    Introduce $Category_n$;
    $k \leftarrow 0$
    $c \leftarrow 1$
    repeat
        Evaluate and correct classifiers by presenting a previously
        unseen instance of $Category_c$
        if $c = n$ then
            $c \leftarrow 1$
        else
            $c \leftarrow c + 1$
        end if
        $k \leftarrow k + 1$
        $p \leftarrow$ precision in last $3n$ question/correction iterations
    until (($p >$ precision threshold and $k \geq n$) or (user sees no
        improvement in precision))
until (user sees no improvement in precision)
```

performs either *teach*, *ask* or *correct* actions (see Algorithm 2[4]). This protocol is exhaustive, generic and applicable to any online, incremental, and open-ended category learning system. Using this protocol, the learning behavior can be followed at each step of its evolution. Classification precision is used as the primary measure. This is an external measure used to control the application of the teaching protocol. For clarity of presentation, the key aspects of the protocol are directly taken from (Seabra Lopes and Chauhan 2008) and repeated here:

For every new category, the instructor introduces, the average precision of the whole system is calculated by performing classification with all known categories. To that end, the instructor repeatedly shows instances of the known categories, checks the agent's predictions and sends corrections when necessary (this is referred below as a "question/correction iteration"). Average precision is calculated over the last $3n$ classification results (n being the number of categories that have already been introduced). The precision of a single classification is either one (correct category) or zero (wrong category). When the number of classification results since the last time a new category was introduced, k, is greater or equal to n but less than $3n$, the average of all results is used. The criterion that indicates that the system is ready to accept a new object category is based on the precision threshold.

The classification precision measure is used to analyze the impact of the introduction of a new category on the learning system, from a possible initial instability to recovery. According to the protocol, a new category is introduced when the precision measure reaches a value above a certain threshold (67% in the present implementation) and at least one instance of each category has been tested. Once an experiment is finished, learning performance is evaluated using two measures: global precision and average classification precision. Global precision is the percentage of correct predictions made during the whole experiment. Average classification precision is the average of all classification precision values computed over all the question/correction iterations.

Although the system is designed to be open-ended, for the reported experiments, the number of categories taught to the robot was set to 13 (8 toys, 3 regular cutlery, and 2 office objects). Two experiments were carried out in which categories were introduced in different sequences. The robot began with zero knowledge and gradually built its vocabulary. In the first experiment, categories were introduced in the following sequence:

cup—4 different objects in this category
star—a star shaped toy
jeep—a toy jeep
scissor—a toy scissor
car—a toy car
horse—a toy horse
fork—2 different objects
stapler—2 different objects
knife—1 object
train—a toy train
bike—a toy bike
boy—a toy in the shape of a small boy
screwdriver—2 different objects

For the second experiment, categories were introduced in the following sequence:

cup, star, jeep, horse, car, bike, fork, knife, stapler, screwdriver, scissor, boy, train.

It is important to note that there was no control over how the user chose to show objects to the robot. To give the reader an idea of how various objects were shown, Fig. 4 provides an illustration. The visual features from the same object can vary significantly depending on how the user decides to demonstrate that object to the robot. In a similar manner, different objects belonging to the same category also lead to very different visual features. However, the learning model of the robot is designed such that the

[4] The description of the "teaching protocol" given here is more detailed than the one in Seabra Lopes and Chauhan (2007, 2008). However, it should be noted that the protocol is still exactly the same. The changes to the algorithmic representation of the protocol were made in order to provide a clearer description.

Table 1 Summary of experiments

Exp#	Precision threshold	# Iterations	# Presented categories	Global precision (%)	Avg. class. precision (%)
1	66.67	148	13	68.24	75.35
2	66.67	245	13	63.27	60.14

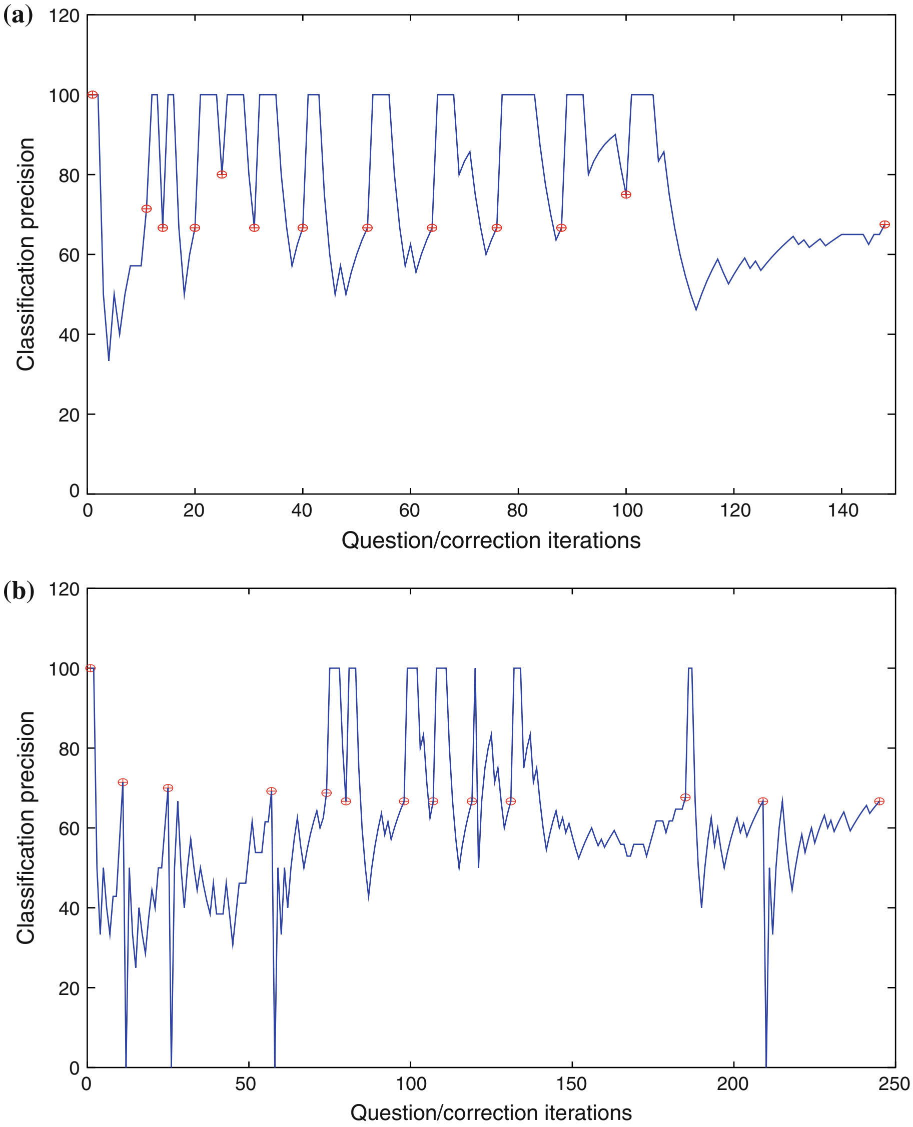

Fig. 5 Evolution of classification precision versus number of question/correction iterations for a precision threshold of 66.67%: **a** experiment 1; and **b** experiment 2. The "circular markers" in the graphs highlight the iteration at which, after the introduction of a new category, the precision threshold was achieved

spoken words guide object category formation. The varying visual features will aggregate together to form a single object category description if the agent identifies that they share the same name (word category).

The agent was able to learn the 13 names and respective meanings in both experiments. Figure 5 presents the evolution of classification precision and Table 1 provides a summary of these experiments. The introduction of a new category normally led to deterioration in classification precision, followed by a period of gradual recovery. In general, the introduction of a new category causes confusion in the prediction of existing categories, hence reducing the classification precision of the system. Each incorrect prediction will lead the user to send a correction. This process is continued till the classification precision reaches the precision threshold (unless it is already above the threshold).The points where the values in the graph are either 100 or 0 are an indicator that after the introduction of a new category (following the teaching protocol), the first category prediction was either correct or incorrect.

Although in both experiments the robot was able to ground the word categories in their respective object categories, as the number of categories taught to the robot increased it became harder for the agent to successfully associate word categories to object categories. This can be noticed for both experiments, where, toward the end of their respective graphs, the interval between the introduction of new categories increases. The length of such interval is an indicator of the amount of effort spent by the agent in reorganizing its categories.

Comparing the two experiments, it is visible that in the second experiment the agent spent much more time in the recovery process. The number of question/correction iterations required to learn 13 categories in the second experiment was 245, which is almost twice as that for the first experiment. In the second experiment, after the introduction of the second category, the agent begins to show difficulty in forming correct category descriptions. The same pattern can be noticed after the introduction of the third, fourth, twelfth, and thirteenth category.

Since the global precision and the average classification precision measures are computed as an average over all the iterations, they reflect the learning performance over a complete experiment. Since the agent had more difficulties in learning categories in the second experiment, the values of these measures for the second experiment are lower. On the other hand, for the first experiment, over all the iterations, classification precision remained mostly above the precision threshold. Therefore, both evaluation measures are higher for the first experiment than those for the second experiment (see Table 1).

The main reason behind the difference in the performances of these experiments is linked to the incremental nature of the learning algorithm. For incremental machine learning algorithms, the order in which new data gets introduced has a great impact on the evolution of the learning performance. For both experiments, there were no constraints over how the user introduced new visual and vocal signals to the agent. The only external control was the order in which new categories can be introduced and the rules of the teaching protocol. We believe, the sequence in which the learning agent received the visual and vocal signals in the first experiment helped the agent in making better word and object category descriptions, early on. However, it is evident from both experiments that the agent was able to learn incrementally and reach the precision threshold for each of the introduced categories.

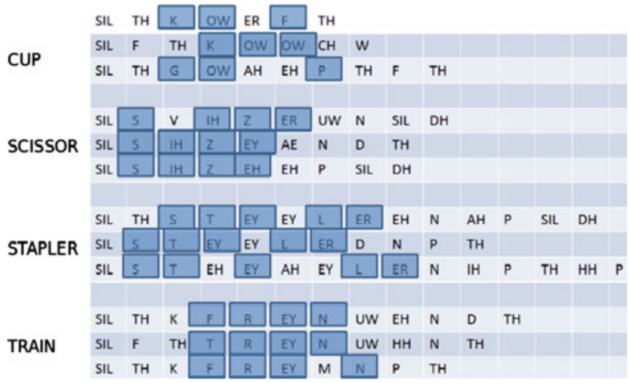
CUP
SIL TH K OW ER F TH
SIL F TH K OW OW CH W
SIL TH G OW AH EH P TH F TH
SCISSOR
SIL S V IH Z ER UW N SIL DH
SIL S IH Z EY AE N D TH
SIL S IH Z EH EH P SIL DH
STAPLER
SIL TH S T EY EY L ER EH N AH P SIL DH
SIL S T EY EY L ER D N P TH
SIL S T EH EY AH EY L ER N IH P TH HH P
TRAIN
SIL TH K F R EY N UW EH N D TH
SIL F TH T R EY N UW HH N TH
SIL TH K F R EY M N P TH

Fig. 6 Phoneme sequences predicted in 3 separate utterances of words *cup*, *scissor*, *stapler*, and *train*, respectively. Highlighted phoneme sequence for each utterance represents the ideal prediction. The non-highlighted phonemes constitute noise in a predicted sequence

Overall, global precision in both experiments is fairly low. This implies that, although the agent shows learning, in the current state its discrimination capability is limited.

A comparative analysis can be made to understand the reason for poor system performance. As mentioned earlier, the visual object classification method used in the present system is derived from our previous work (Chauhan and Seabra Lopes 2010; Seabra Lopes and Chauhan 2008). Our approach to vocabulary grounding involved incrementally teaching meanings of object names typed directly into the computer (it was possible to learn up to a total of 69 object categories). Category name prediction was achieved by direct classification of objects. In the present scenario, the use of spoken words adds extra complexity to the vocabulary grounding problem. As a perceptual input, a spoken word (audio signal) is uncertain. Classifying a spoken word as belonging to a word category relies on the measure of similarity (*S_word*) presented in "Word similarity and classification". A key component in computing this measure is the predicted phoneme sequence for a spoken utterance. However, the phoneme sequences predicted for the spoken utterances of the same word can vary a lot (see Fig. 6). Such predictions add a significant amount of noise to a word representation and, more generally, to word categories.

5.2 Evaluation of the word similarity measure

To perform an independent evaluation of the similarity measure, we created a dataset of isolated spoken words. The words chosen for the evaluation were the ones used in the previous two experiments. This dataset was collected with the help of 8 volunteers (5 men and 3 women), none of whom had English as their first language. Age of the participants ranged between 28 and 33. A small software program was developed which flashed the name of the word to be spoken and the participants were given a 2 seconds time-frame to speak each word. In total, 5 utterances per word per person were recorded. This led to a collection of 520 (5 × 13 × 8) isolated spoken words. Simultaneously, a text file was generated which contained the name (in text) and the location of each of the stored audio files.

We used 10-fold cross-validation to evaluate the similarity measure nearest-neighbor classification scenario. The dataset was evenly divided into 10 subsets of equal size, such that each subset contained equal number of instances of each of the 13 words. For each validation step, one of the subsets was taken out as the test/validation set and the word categories were represented by all the remaining instances. In each validation step, each instance in the test set was classified according to the nearest-neighbor algorithm. An average classification precision of 74.81% (±6.81) was achieved.

The similarity measure was compared with a standard distance metric (Levenshtein distance), normally used in string matching algorithms (Navarro 2001). Using this metric, the distance between two strings is computed as the minimum number of edit operations (insertion, deletion, and substitution actions) required to transform one string into another. To use this metric, we only used the phoneme sequence to represent a word, discarding the MFC set. Using 10-fold cross-validation, the average classification precision of 80.96% (±6.81) was achieved.

These results indicate that, for comparing word representations, finding the cheapest path (Levenshtein distance) gives a better performance than following the path centered around the diagonal elements (our method). Although the difference in performances is not huge, considering that Levenshtein distance did not use any extra information (e.g., the MFC set), it can be a cheaper and most likely a better alternative to our method. In the future, we plan to qualitatively investigate different types of word representations and distance metrics within the framework of the learning architecture presented in this paper.

6 Discussion and conclusions

In this paper, we presented a novel strategy for dynamic category formation in artificial cognitive agents, where spoken words assist in building object categories. This strategy has its inspiration in early child language development literature, where studies have shown a strong influence of naming on conceptual development. The key assumption here is that word categories (especially common nouns) are homogeneous in comparison to their corresponding object categories. The objective of the approach outlined here was to exploit the homogeneity in utterances, to gradually form new or improve existing word categories as new word instances become available. A word category has one-to-one correspondence with its corresponding object category and all the changes to word category descriptions get transmitted to their corresponding object categories as well.

We developed a physically embodied agent with a camera and a microphone for visual and auditory perception. A social language grounding scenario was designed, where a human instructor taught the agent the names of the objects present in their shared visual environment.

The MFC set and the phoneme sequence were used to describe a word. A simple instance-based approach was used for word category representation, where a set of words describes the word category. Similarly, an object category is described by a set of representations of instances of that category. A word category is grounded in its object category such that each word in the word category is coupled with an instance in the corresponding object category.

A new membership measure based on DTW and a local greedy search algorithm was developed for categorization of the speech input. For two word representations being compared, this algorithm finds a locally optimal path which is centered around the diagonal elements of the *DTW* matrix.

An approach was also developed for clustering instances in a word category, leading to either transfer of clusters between categories or creation of new categories. The clustering approach is critical for the methodology developed for object category formation and organization, which takes the word category representations into account for dynamically organizing object categories. This is a key contribution of the paper. Another contribution of the paper, is the open-ended nature of the language acquisition process. For each *teach* or *correct* action from the human user, this algorithm is called and relevant object category representations are modified/updated/reorganized, so as to provide better categorization.

To evaluate the system on the task of vocabulary acquisition, two experiments were conducted using a clearly defined protocol. Although the learning architecture of the agent is open-ended, for these experiments, a limit of 13 words was set on the vocabulary to be taught. The results are encouraging and, we believe, the approach is both interesting and viable, but further improvements are required.

To perform an independent evaluation of our word similarity measure, we collected a dataset of 520 word utterances. We compared the metric proposed in this paper with Levenshtein distance. The results lead to the conclusion that locating the cheapest path (Levenshtein distance) gives a better performance than following the path centered around the diagonal elements (our method).

In the approach described in this paper, one form of perceptual input (spoken words) guides the organization of another form of perceptual input (visual) to build object categories. Does that mean that the inverse is also possible? Using the same approach, could visual input guide the formation of word categories? As mentioned earlier, the key assumption of our work is that word categories are homogeneous. This homogeneity is exploited to cluster the most similar word instances and form/reorganize word categories (and modify the associated object categories). For object categories, this type of homogeneity is not very common. For example, the word *bottle* refers to an object category whose instances can vary greatly from one object to another. In this example, the lack of homogeneity in the visual domain would lead to the formation of separate clusters. These clusters would form separate object categories and their associated word instances would form separate word categories. In other words, instances of the same word would be spread through multiple visual categories, leading to confusion.

As a final note, it is important to state here that object categories can be (and are) learned without the need of a language. But there is a mounting evidence, in humans, language (especially words) plays a big role in facilitating conceptual development. The presented architecture is designed to test whether, only using the information contained in vocal symbols, visual categories can be learned by a robotic agent. The results reported here lead to the conclusion that spoken words can be successfully used for learning visual object categories.

Acknowledgments The authors would like to thank the 8 participants, who helped in the creation of isolated spoken word dataset. The authors also wish to thank the anonymous reviewers for constructive comments and suggestions for improving the quality of the manuscript.

References

Ballard DH, Yu C (2003) A multimodal learning interface for word acquisition. In: International conference on acoustics, speech and signal processing

Barsalou L (1999) Perceptual symbol systems. Behav Brain Sci 22(4):577–609

Bates E, Thal D, Finlay B, Clancy B (1992) Early language development and its neural correlates. In: Segalowitz SJ, Rapin I (eds) Elsevier, Amsterdam, pp 69–110

Bloom P (2000) How children learn the meanings of words. MIT Press, Cambridge

Bloom P (2001) Word learning. Curr Biol 11:5–6

Bomba PC, Siqueland ER (1983) The nature and structure of infant form categories. J Exp Child Psychol 37:609–636

Cangelosi A, Harnad S (2000) The adaptive advantage of symbolic theft over sensorimotor toil: grounding language in perceptual categories. Evol Commun 4(1):17–142

Chauhan A, Seabra Lopes L (2010) Acquiring vocabulary through human robot interaction: a learning architecture for grounding words with multiple meanings. In: AAAI Fall symposium on Dialog with Robots (DWR2010). Arlington, VA

Cowley SJ (2007) Distributed language: biomechanics, functions and the origins of talk. In: Lyon C, Nehaniv C, Cangelosi A (eds) Springer, Berlin, pp 105–109

Crystal D (1987) How many words? Engl Today 12:11–14

Fenson L, Dale PS, Reznick JS, Bates E, Thal D, Pethick S (1994) Variability in early communicative development. Monogr Soc Res Child Dev 59(5)

Gillette J, Gleitman H, Gleitman L, Lederer A (1999) Human simulations of vocabulary learning. Cognition 73:135–176

Gold K, Doniec M, Crick C, Scassellati B (2009) Robotic vocabulary building using extension inference and implicit contrast. Artif Intell 173(1):145–166

Grauman K, Darrell T (2007) The pyramid match kernel: efficient learning with sets of features. J Mach Learn Res 8:725–760

Harnad S (1990) The symbol grounding problem. Physica D 42:335–346

Jusczyk PW (1993) From general to language specific capacities: the wraspa model of how speech perception develops. J Phonet 21:3–28

Jusczyk PW, Aslin RN (1995) Infants detection of the sound patterns of words in fluent speech. Cogn Psychol 29:1–23

Krunic V, Salvi G, Bernardino A, Montesano L, Santos-Victor J (2009) Affordance based word-to-meaning association. In: IEEE International conference on robotics and automation. Kobe, Japan

Landau B, Smith LB, Jones S (1988) The importance of shape in early lexical learning. Cogn Dev 3:299–321

Levinson SE, Squire K, Lin RS, McClain M (2005) Automatic language acquisition by an autonomous robot. In: AAAI Spring symposium on developmental robotics

Loreto V, Steels L (2007) Social dynamics: emergence of language. Nat Phys 3:758–760

Love N (2004) Cognition and the language myth. Lang Sci 26(6): 525–544

Messer DJ (1994) The development of communication. Wiley, Chichester

Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv 33(1):31–88

Rath M, Manmatha R (2003) Word image matching using dynamic time warping. In: Computer vision and pattern recognition (CVPR)

Roy D (2003) Grounded spoken language acquisition: experiments in word learning. IEEE Trans Multi 5(2):197–209

Roy D, Pentland A (2002) Learning words from sights and sounds: a computational model. Cogn Sci 26:113–146

Seabra Lopes L, Chauhan A (2008) Open-ended category learning for language acquisition. Connect Sci 20(4):277–297

Seabra Lopes L, Chauhan A (2007) How many words can my robot learn? an approach and experiments with one-class learning. Interact Stud 8(1):53–81

Seabra Lopes L, Chauhan A, Silva J (2007) Towards long-term visual learning of object categories in human-robot interaction. In: Neves JM, Santos M, Machado J (eds) New trends in artificial intelligence. Ass. Portuguesa para a Inteligłncia Artificial. pp. 623–634.

Skočaj D, Berginc G, Ridge B, Štimec A, Jogan M, Vanek O, Leonardis A, Hutter M, Hawes N (2007) A system for continuous learning of visual concepts. In: International conference on computer vision systems ICVS 2007. Bielefeld, Germany

Smith LB, Samuelson L (2006) An attentional learning account of the shape bias: reply to cimpian and markman (2005) and booth, waxman, and huang (2005). Dev Psychol 42(6):1339–1343

Steels L (2008) The symbol grounding problem has been solved. so what's next?. In: Vega M (eds) Symbols and embodiment: debates on meaning and cognition, vol 12. Oxford University Press, Oxford

Steels L, Kaplan F (2002) AIBO's first words: the social learning of language and meaning. Evol Commun 4(1):3–32

Thomaz AL, Breazeal C (2008) Teachable robots: understanding human teaching behavior to build more effective robot learners. Artif Intell J 172:716–737

Thomaz AL, Hoffman G, Breazeal C (2006) Experiments in socially guided machine learning: Understanding how humans teach. In: Proceedings of the 1st annual conference on human-robot interaction (HRI)

Waxman SR (2008) All in good time: how do infants discover distinct types of words and map them to distinct kinds of meaning? Infant pathways to language: methods, models, and research directions. In: Colombo J, McCardle P, Freund L (eds) pp 99–118

Yeung HH, Werker JF (2009) Learning words' sounds before learning how words sound: 9-month-old infants use distinct objects as cues to categorize speech information. Cognition 113(2):234–243

Yoshida H, Smith LB (2005) Linguistic cues enhance the learning of perceptual cues. Psychol Sci 16(2):90–95

Yu C, Ballard DH (2004) A multimodal learning interface for grounding spoken language in sensory perceptions. ACM Trans Appl Percept 1:57–80

Yu C, Ballard DH (2007) A unified model of early word learning: Integrating statistical and social cues. Neurocomputing 70(13–15):2149–2165