Perceptually Grounded Word Meaning Acquisition: A Computational Model

Claudius Gläser, Frank Joublin

2010

Preprint:

This is an accepted article published in Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci). The final authenticated version is available online at: https://doi.org/[DOI not available]

Perceptually Grounded Word Meaning Acquisition: A Computational Model

Claudius Gläser (claudius.glaeser@honda-ri.de)
Honda Research Institute Europe
Carl-Legien-Strasse 30, 63073 Offenbach, Germany

Frank Joublin (frank.joublin@honda-ri.de)
Honda Research Institute Europe
Carl-Legien-Strasse 30, 63073 Offenbach, Germany

Abstract

We present a computational model for the incremental acquisition of word meanings. Inspired by Complementary Learning Systems theory the model comprises different components which are specifically tailored to satisfy the contradictory needs of (1) rapid memorization of word-scene associations and (2) statistical feature extraction to reveal word meanings. Both components are recurrently coupled to achieve a memory consolidation. This process reflects itself in a gradual transfer of the knowledge about a word's meaning into the extracted features. Thereby, the internal representation of a word becomes more efficient and robust. We present simulation results for a visual scene description task in which words describing the relations between objects have been trained. This includes relations in size, color, and position. The results demonstrate our model's capability to acquire word meanings from few training exemplars. We further show that the model correctly extracts word meaning-relevant features and therefore perceptually grounds the words.

Keywords: Word Learning; Computational Model; Complementary Learning Systems; Categorization

Introduction

When hearing a novel word, a language learner has to associate the word with its meaning. Establishing such word-meaning mappings is an inherently difficult task as the learner initially cannot know to what the word refers to. Quine (1960) illustrated this problem with the example of a stranger who hears a native saying "gavagai" after seeing a rabbit. How can the stranger determine the meaning of "gavagai"? It may refer to the rabbit, a part of the rabbit, its color, any fast moving animal, or even that a rabbit is tasty. This problem, usually referred to as *referential uncertainty*, cannot be solved from a single word-scene pairing. Rather the use of the word in different contexts enables the learner to extract its meaning. Nevertheless, children learn the meaning of words from few exposures to them. They rapidly construct hypotheses about word meanings, which may initially be linked to specific contexts in which the words occurred. Over time, however, children generalize among different observations, even though this may result in an overextension of a word's use (MacWhinney, 1998). This remarkable ability of children has been subject to many studies and resulted in numerous theories on early word learning.

In this paper we present a computational model for the incremental acquisition of word meanings which is inspired by the learning capabilities of children. More precisely, the system has been designed to rapidly build internal representations of words from few training samples. The thus acquired knowledge can be used to generalize to previously unseen scenes. Moreover, the framework is endowed with a learning mechanism that extracts features which are relevant to the core meaning of a word. This is done by exploiting the statistical evidence which resides from a word's use in different contexts. Our model tightly couples the rapid memorization of word-scene associations with the statistical feature extraction. This results in learning dynamics which resemble a gradual knowledge transfer and consolidation.

We will present experimental results which validate the model. Therefore, the model has been applied in a simulated visual scene description task where words for the relations between pairs of geometric objects have been trained. This includes relations in position, color, and size. The results from this experiment illustrate that our model rapidly acquires word meanings from few training exemplars and further extracts word meaning-relevant features.

The remainder of this paper is organized as follows. Next, we will review existing approaches for word meaning acquisition and relate our model to them. Afterwards, we will state contradictory needs that computational models have to satisfy. We proceed with the presentation of our computational model and subsequently show experimental results for it. Finally, we give a summary and outline our future work.

Related Work

Existing computational models address different levels of referential uncertainty. Firstly, there are approaches which consider the problem of how a learner establishes a mapping between words and a set of pre-defined meanings (e.g. Siskind, 1996; K. Smith, Smith, Blythe, & Vogt, 2006; Fontanari, Tikhanoff, Cangelosi, Ilin, & Perlovsky, 2009). In these models the first occurrence of a word typically induces multiple hypotheses about its meaning. These hypotheses become subsequently pruned either by incorporating learning constraints (Markman, 1990) or via *cross-situational learning* (L. Smith & Yu, 2008) - a technique making use of the statistical evidence across many individually ambiguous word-scene pairings. However, these models disregard the fact that learners can seldom rely on a set of pre-established concepts. Word meanings rather become flexibly constructed and shaped through language use (Boroditsky, 2001).

Therefore, a second group of models further asks how language use yields sensori-motor concepts to which words become associated (e.g. Steels & Kaplan, 2002; Skocaj et al., 2007; Kirstein, Wersing, Gross, & Körner, 2008; Wellens, Loetzsch, & Steels, 2008). In these models the learner observes the world through multiple (analog or discretized) input channels. The words finally serve as labels for categories, which become incrementally constructed on the multi-dimensional input space and gradually refined by concentrating on the most important input dimensions.

Lastly, there are models which aim at the acquisition of both phonological form and semantic form. Such models either build perceptual clusters in the acoustic space and the semantic space and subsequently associate them with each other (Yu & Ballard, 2003; Goerick et al., 2009) or clustering is directly carried out in the joint acoustic-semantic space (Roy & Pentland, 2002).

The model we present in this paper falls into the second group of methods, i.e. based on the observation of multiple word-scene pairs it acquires perceptual categories by which the words become grounded. To achieve realistic word meaning acquisition we further place several requirements on our model: (1) It should be capable of learning during online operation. Consequently, the model has to apply incremental learning techniques as training exemplars sequentially arise during a learner's interaction with its environment. (2) The model should further rapidly learn from few examples and afterwards apply the acquired knowledge to generalize to novel scenes. (3) However, to be efficient and robust the internally built categories should reflect the core structure underlying the word meanings. Thereby, we use the term core structure to refer to the essential aspects which define the meaning of a word. (4) Lastly, for systems with minimum predefined knowledge this core structure is usually hidden and thus cannot be directly accessed by concentrating on input dimensions which carry the meaning. The model rather has to extract word meaning-relevant feature dimensions in terms of a transformation from the input space.

The combination of these requirements is what distinguishes our model from existing approaches. Particularly the combination of rapid incremental learning with word meaning-relevant feature extraction has (to our best knowledge) not been realized previously. In (Skocaj et al., 2007; Wellens et al., 2008) and most notably (Kirstein et al., 2008) feature selection is applied, i.e. the learning focuses on the input dimensions which are considered to be relevant for representing the word meanings. By doing so the approaches inherently rely on the assumption that words can be grounded in a subset of the input dimensions. This in turn means that significant knowledge about the words to learn has to be put into the system by the designer. In contrast, our system generates new word meaning-relevant feature dimensions out of a set of basic input dimensions. We consider this ability to be crucial for life-long incremental learning systems for which the extent of words to be learned is unknown at design time.

Complementary Learning Systems Theory

The way how children acquire the meaning of new words is fascinating in multiple respects. When they hear a word for the first time they already get a glimpse on what it may mean. This ability may be facilitated by learning constraints or biases (Markman, 1990). It is anyway non-disputable that even the exposure to just a few uses of the word enables the child to generalize and apply the word in novel contexts. Even though generalization may occasionally result in errors (MacWhinney, 1998), over time children robustly identify the core meaning of a word.

A Computational Learning Dilemma

Modeling word meaning acquisition computationally, however, is difficult as contradictory needs have to be simultaneously satisfied. McClelland, McNaughton, and O'Reilly (1995) illustrated this fact on the example of artificial neural network models: On the one hand, the learning from few training samples requires a rapid or even one-shot memorization of the items which can be achieved by using high learning rates. This implies that localized representations, which keep the memory items separated from each other, have to be used. Otherwise, a neural network would suffer from *catastrophic forgetting* - the problem that the incorporation of new knowledge overwrites previously memorized items. On the other hand, the extraction of the core structure underlying a word meaning necessitates a statistical learning approach as knowledge has to be accumulated over many training exemplars. Such a learning can be achieved using low learning rates and overlapping representations. Artificial systems, which learn from few examples while they simultaneously extract statistical evidence, are consequently difficult to achieve.

A Solution to the Problem

Obviously, humans (and particularly children) successfully solve this learning task. Endowing artificial systems with mechanisms inspired by human learning may consequently lead a way to overcome the dilemma. *Complementary Learning Systems (CLS) Theory* (McClelland et al., 1995) suggests that the human brain makes use of separate but tightly coupled learning and memory devices which are specifically tailored to satisfy the contradictory needs. More precisely, it is proposed that new memories are first stored in the hippocampal system which is known to perform rapid learning while utilizing localized representations. The hippocampal system further allows the reactivation of recent memories during rest or sleep. This reactivation in turn enables neocortical areas to extract the core structure underlying different memories via *interleaved learning* - a technique where new items become gradually learned while learning is interleaved with the memorization of other items. Consequently, a gradual memory consolidation and transfer from the hippocampal system to neocortical sites can be observed. Furthermore, there is behavioral and neuroscientific evidence which is in accordance with a CLS theory for the lexical and semantic acquisition of novel words (Davis & Gaskell, 2009).

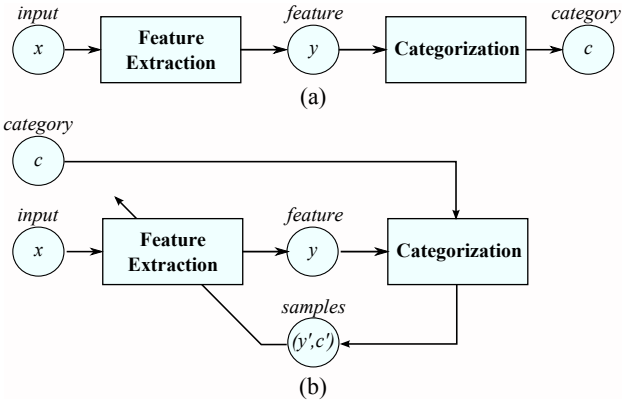


Figure 1: Architecture of the computational model: (a) Input samples x become transformed into feature patterns y which are subsequently categorized. (b) During learning the system components are recurrently coupled (see text for details).

Computational Model

In what follows we treat word learning as category learning. This is reasonable as a word refers to collections of entities which belong to the same category. Word meanings are consequently the conditions underlying category membership (Bloom, 2000). We restrict our description to the learning of one word. Multiple words can be learned straightforwardly by creating multiple instances of the system. As shown in Fig. 1, the framework consist of a feature extraction layer and a categorization layer which are recurrently coupled. The feature extraction transforms an input pattern x into a feature pattern y for which a category membership c is subsequently calculated. Here, c is a binary variable which signals whether the category's word label is appropriate for the description of the input pattern ($c=+1$) or not ($c=-1$).

Our model is largely inspired by CLS theory. Nevertheless, the model is not meant to provide a 1:1 mapping to certain brain areas. It rather resembles CLS theory from a functional perspective. For this reason, we will highlight functional correspondences of our model with different brain areas.

Feature Extraction

In the feature extraction layer word meaning-relevant features, which facilitate the subsequent categorization of a pattern, should become extracted. The learning consequently has to exploit the statistical evidence stemming from the observation of multiple word-scene pairings. Such a statistical feature extraction is obviously part of neocortical learning.

In (Hild, Erdogmus, Torkkola, & Principe, 2006) a learning technique called Maximizing Renyi's Mutual Information (MRMI) has been proposed. MRMI tries to maximize the information that the feature patterns carry about category memberships. Hence it is ideally suited to accomplish the learning task. We restrict learning to a linear feature extraction of form $y = R \cdot x$. We consequently aim at the identification of a transformation matrix R such that the mutual information $I(Y;C) = H(Y) - H(Y|C)$ between the feature patterns and

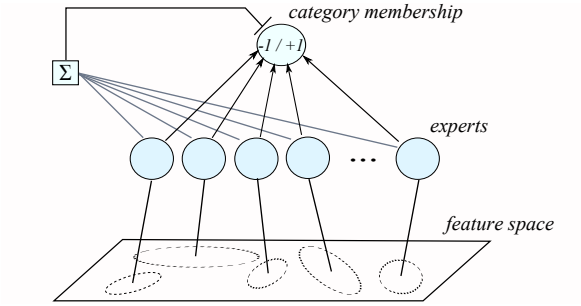


Figure 2: The architecture of an NGnet.

category labels becomes maximized. By relying on Renyi's quadratic entropy $H_2(Y)$ and its estimation using Parzen windows (Hild et al., 2006) the criterion to be maximized is

$$I(Y;C) = -\log \frac{1}{K} \sum_{k=1}^{K} G(y(k) - y(k-1), 2\sigma^2) + \sum_{j \in \{-1,+1\}} \left(\frac{K_j}{K} \log \frac{1}{K_j} \sum_{k=1}^{K_j} G(y_j(k) - y_j(k-1), 2\sigma^2) \right). \quad (1)$$

Here, $G(z, \sigma^2 I) = \exp(-\frac{1}{2} \frac{z^T z}{2\sigma^2})$ is a Gaussian kernel, $y_{+1}(k)$ and $y_{-1}(k)$ denote the k-th exemplars of feature patterns belonging to a category or not, K_{+1} and K_{-1} are the numbers of such patterns, and $K = K_{+1} + K_{-1}$. Since $y(k) = R \cdot x(k)$, we can estimate R via stochastic gradient ascent on $I(Y;C)$.

To de-correlate the feature dimensions and to perform dimensionality reduction we additionally apply Principal Component Analysis (PCA) on the extracted features. By assuming the inputs x to be white with zero mean and unit variance, the principal feature dimensions can be obtained via eigendecomposition of $R \cdot R^T$. Let Ψ be the matrix of eigenvectors whose cumulative energy content exceeds a threshold. Then we calculate feature patterns y according to

$$y = \Omega \cdot x = \Psi^T \cdot R \cdot x. \quad (2)$$

Categorization

To incrementally learn a category we use an adaptive Normalized Gaussian Network (NGnet) which we recently proposed (Gläser & Joublin, 2010). As shown in Fig. 2, the NGnet is composed of multiple locally operating experts, each of them being responsible for features stemming from its associated input region. The category membership $c \in \{-1, 1\}$ of a feature pattern y is calculated according to

$$c(y) = \text{sign} \left[\frac{1}{\sum_{j=1}^{M} \phi_j(y)} \cdot \sum_{i=1}^{M} \alpha_i \cdot \phi_i(y) \right]. \quad (3)$$

Here, M denotes the number of experts and α_i the weight of expert i to the output neuron. Furthermore, $\phi_i(y)$ is the response of the i-th expert to feature y which is described by a multivariate Gaussian of form

$$\phi_i(y) = \exp \left(-\frac{1}{2} \cdot (y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i) \right), \quad (4)$$

where μ_i and Σ_i denote the center and covariance matrix of the Gaussian. The decision whether a feature pattern belongs to a category is finally obtained by application of the sign function to the continuously valued output. The network parameters are determined during online operation via Expectation-Maximization (EM) training as proposed in (Xu, 1998).

Since the NGnet statistically learns an internal category representation which associates inputs from different modalities, our categorization layer functionally resembles multimodal associative cortices, e.g. the perirhinal cortex. However, our adaptive NGnet is additionally endowed with mechanisms which allow a demand-driven allocation and removal of experts (Gläser & Joublin, 2010). This enables the network to perform a one-shot memorization of word-scene associations. Our categorization layer consequently also models the rapid initial learning as it is carried out by the hippocampus.

More precisely, our model accomplishes network growth and pruning as follows: (1) New word-scene associations become memorized based on the novelty or surprise of an input sample. Similarly, already memorized associations become (2) pruned if they became redundant, (3) split if the internal representation has to be refined, or (4) merged if they are sufficiently similar. For a detailed description of these mechanisms we refer to (Gläser & Joublin, 2010).

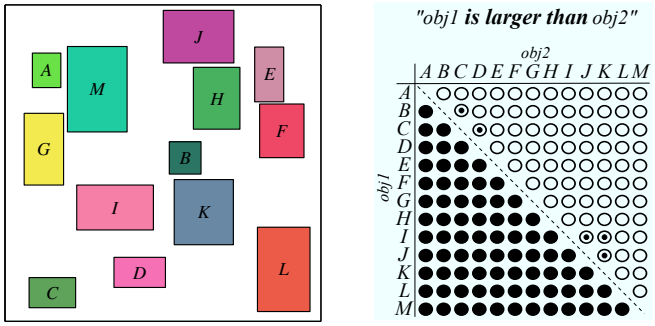
Coupling of the Components

Inspired by CLS theory we finally couple the slow statistical feature extraction and the rapid category learning. As shown by the pseudo-code in Alg. 1 the incremental learning mechanism consists of four steps which are carried out every time a new training exemplar is obtained.

Algorithm 1 Incremental Learning

```
Initialize the feature extraction to R = I, Ψ = I
Initialize an empty NGnet
for all training samples (x, c) do
    Update the NGnet with (y, c)
    Generate a set of samples (y', c') using the NGnet
    Train the feature extraction on the generated samples
    Adapt the NGnet to the changed feature space
end for
```

After updating the NGnet with a training sample, the internal representation of a category is used to reactivate memorized associations. This step resembles hippocampal dreaming. We consequently produce a set of samples (y', c') composed of feature patterns y' and associated category memberships c'. To do so, we first determine whether a local expert i represents category members ($c' = +1$) or non-members ($c' = -1$) and next randomly draw feature patterns y' from its Gaussian-shaped receptive field. Since the receptive field is described by its mean μ_i and covariance matrix Σ_i, a feature pattern y' can be generated by $y' = \mu_i + B \cdot z$, where $z \sim \mathcal{N}(0, I)$ is a random vector and B is obtained from the Cholesky decomposition $B \cdot B^T = \Sigma_i$.



(a) (b)

Figure 3: In (a) an example scene used in the visual description task is depicted. In (b) the output of the model after learning the meaning of *is larger than* is shown. Black circles correspond to category members, white circles to non-members, and dotted circles denote errors made by the system.

Afterwards, the generated samples are used to train the feature extraction. In other words, the feature extraction searches for commonalities among the reactivated patterns and tries to extract the condition which discriminates between members and non-members of the category. This learning process changes the feature space the categorization layer is operating on. For this reason, we finally adapt the NGnet to the changed feature space in an analytic way. Since we use a linear feature extraction, the change in the feature space can be expressed in terms of an affine transformation $\tilde{y} = A \cdot y$ with $A = \tilde{\Omega} \cdot \Omega^{-1}$. Here, Ω and $\tilde{\Omega}$ denote the feature extraction matrices before and after the learning. We consequently adjust a local expert's receptive field by calculating its new center $\tilde{\mu}_i$ according to $\tilde{\mu}_i = A \cdot \mu_i$ as well as its associated covariance matrix $\tilde{\Sigma}_i$ according to $\tilde{\Sigma}_i = A \cdot \Sigma_i \cdot A^T$.

Since these learning steps are carried out iteratively, knowledge about a category becomes consolidated as more training samples are processed. The knowledge, which has been first acquired in the categorization layer (via the memorization of word-scene associations), becomes gradually transfered into the extracted features. Due to the fact that the extracted features facilitate the categorization task, this knowledge transfer leads to a more robust categorization as well as a less complex NGnet needed to represent the category.

Experimental Results

We evaluated our computational model in a visual scene description task in which the meaning of words for the relations between objects has to be acquired. Thereby, a learner and a tutor observe a scene composed of geometric objects as the one shown in Fig. 3(a). The tutor selects two out of the objects and describes the relation between them, e.g. by saying

"K is larger than D."

Based on such exemplars of word use the learner has to incrementally build-up internal concepts which correspond to the words' meanings. The training of the model is illustrated in

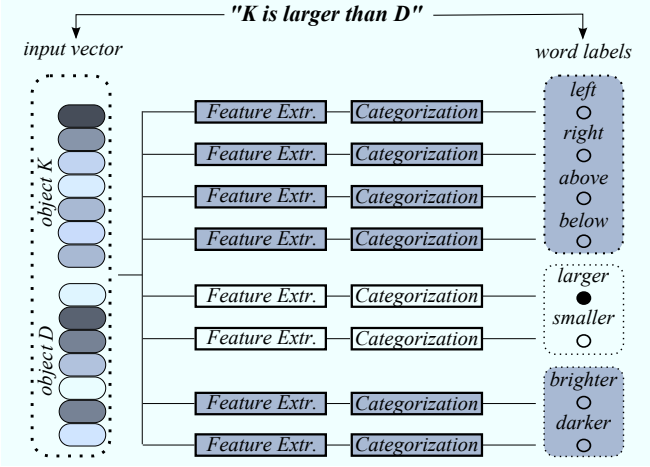


Figure 4: The training of the model in the visual scene description task is illustrated (see text for details).

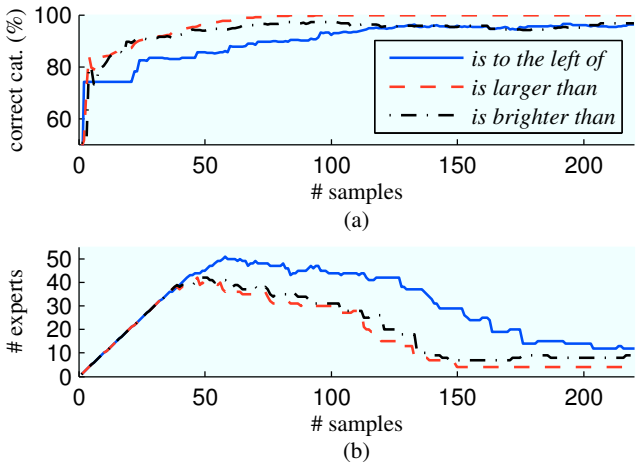


Figure 5: The evolution of (a) the correct categorization rate and (b) the complexity of the NGnet is shown for the learning of different words.

Fig. 4. For the present experiment we consider the learner to have sufficient syntactical knowledge to identify the objects of interest (e.g. *K* and *D*) as well as the word to be learned (e.g. *is larger than*). For computational purposes we further did not carry out the experiment in direct interaction with the system, but rather used simulated scenarios which provide a ground truth for performance evaluation.

Each of the objects in a scene is represented by its absolute position, its width and height, as well as its RGB color value. Consequently, tuples composed of a 14-dimensional perceptual vector (7 dimensions per object) as well as a word label served as training inputs to the system. Words for object relations concerning position (*is to the left of*, *is to the right of*, *is above*, *is below*), size (*is larger than*, *is smaller than*), and color (*is brighter than*, *is darker than*) has been trained. However, it is important to note that the system did not have prior knowledge about the relevance of input dimensions with respect to the meaning of the words. In contrast, important dimensions (e.g. the relative object positions) are even not present and have to be extracted by the system. For each of the words to learn we applied an adaptive NGnet as a binary categorization module and further extracted word meaning relevant features. To cope with missing negative training data we implemented the mutual exclusivity bias between words related to object positions, sizes, and colors, respectively. In other words, a positive training exemplar for *is larger than* has been additionally used as negative training sample for *is smaller than* (Regier, 1996).

The results of this experiment are shown in Fig. 5. In (a) we plot the system performance for the learning of individual words. The performance (correct categorization rate) has been determined on a set of scenes not included in the training data. In (b) we further plot the complexity of the individual classifiers for which the number of local experts comprising the NGnet is an indicator. To keep the plots readable, here we restrict ourselves to curves for the learning of *is to the left of*, *is larger than*, and *is brighter than*. The learning of the other words resulted in qualitatively similar curves.

From the plots we see that the system performance rapidly increases during the presentation of the first training exemplars and afterwards converges towards a near optimal level. In contrast, the complexities of the classifiers also increase at the beginning, but subsequently decrease and maintain a low level afterwards. The observed behavior of the model is in-line with CLS theory, insofar as it can be explained by two complementary learning processes which run at different timescales: (1) Initially, new knowledge is rapidly memorized. In our model this is accomplished by the on-demand allocation of local experts within the classifier. After a while, the experts adequately represent upcoming training samples such that they do not have to be memorized additionally. Consequently, the classifier complexity as well as the system performance increase at the beginning. (2) Afterwards, knowledge is gradually transferred. In our model the knowledge shifts into the iteratively extracted word meaning-relevant features. These features facilitate the classification task such that a less complex classifier can be applied. At the same time, however, the internal representation of a word meaning becomes more robust and, thus, further increases the system performance.

After training, an analysis of the extracted features revealed that the built categories solely rely on the meaning of the corresponding words. For example, for representing the meaning of *is larger than* the feature

$$(width_{obj_1} + height_{obj_1}) - (width_{obj_2} + height_{obj_2})$$

has been extracted which is an adequate linear approximation of the real decision criteria

$$(width_{obj_1} \cdot height_{obj_1}) - (width_{obj_2} \cdot height_{obj_2}) > 0.$$

Similarly, the relative horizontal and vertical positions have been extracted for the description of spatial relations. This shows that our framework is able to acquire the meaning of words and consequently grounds them.

Finally, the output of the classifiers can be used to describe a visual scene. For the scenario depicted in Fig. 3(a), we show the output of our framework concerning the judgment whether an object *is larger than* another object in Fig. 3(b). As can be seen, objects pairs are correctly categorized except for rare cases in which the object sizes are very similar.

Summary & Future Work

In this paper we presented a computational model for the incremental acquisition of word meanings. The novelty of the framework stems from its combined ability to (1) rapidly build categories which correspond to the learned words while (2) it simultaneously extracts features which underly the meaning of the words. We consider these abilities to be fundamental for life-long incremental learning systems which have to cope with minimal predefined task knowledge. To satisfy the contradictory needs of rapid learning from few examples as well as statistical feature extraction we modeled learning mechanisms which are known to be beneficial for humans. More precisely, our framework resembles CLS theory insofar as it uses separate but tightly coupled components which are specifically tailored to meet these criteria.

We evaluated our model in a visual scene description task, where words for the relations between objects have been taught. Our results demonstrate that the system acquires word meanings based on the observation of just a few word-scene pairings. It subsequently uses its knowledge to generalize to novel scenes. The results further showed that the system implements a memory consolidation process in which knowledge about a word's meaning gradually shifts from the rapidly learned category representation into the slowly extracted features. This consolidation process is beneficial as it abstracts the core meaning of a word and, hence, lets the internal representation of a word become more robust and efficient.

Part of our future work will be the extension of the model to incorporate a non-linear feature extraction. This would allow the system to extract more complex dependencies which may underly a word's meaning. Secondly, we will endow the model with a mechanism which detects the mutual exclusivity between words. This learning bias is currently pre-defined, but has to be autonomously applied by the system to enable a learning of an arbitrary set of words. Lastly, we will extend our teaching scenario to include social learning. Social learning enables an active learning by the system which is useful for testing hypotheses about a word's meaning.

References

Bloom, P. (2000). *How children learn the meaning of words*. MIT Press.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1–22.

Davis, M., & Gaskell, M. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Phil. Trans. R. Soc. B*, *364*(1536), 3773–3800.

Fontanari, J., Tikhanoff, V., Cangelosi, A., Ilin, R., & Perlovsky, L. (2009). Cross-situational learning of object-word mapping using Neural Modeling Fields. *Neural Networks*, *22*(5-6), 579–585.

Gläser, C., & Joublin, F. (2010). An adaptive normalized gaussian network and its application to online category learning. In *Proc. IJCNN.*

Goerick, C., Schmuedderich, J., Bolder, B., Janssen, H., Gienger, M., Bendig, A., et al. (2009). Interactive online multimodal association for internal concept building in humanoids. In *Proc. IEEE-RAS Int. Conf. on Humanoids.*

Hild, K., Erdogmus, D., Torkkola, K., & Principe, J. (2006). Feature extraction using information-theoretic learning. *IEEE Trans. on Pattern Anal. and Machine Intell.*, *29*(9), 1385–1392.

Kirstein, S., Wersing, H., Gross, H. M., & Körner, E. (2008). An integrated system for incremental learning of multiple visual categories. In *Proc. ICONIP.*

MacWhinney, B. (1998). Models of the emergence of language. *Annu Rev Psychol*, *49*, 199–227.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev*, *102*(3), 419–457.

Quine, W. V. O. (1960). *Word and object*. MIT Press.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113-146.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Skocaj, D., Berginc, G., Ridge, B., Vanek, O., Hutter, M., & Hawes, N. (2007). A system for continuous learning of visual concepts. In *Proc. ICVS.*

Smith, K., Smith, A. D. M., Blythe, R. A., & Vogt, P. (2006). Cross-situational learning: A mathematical approach. In *Symbol grounding and beyond.* Springer.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Steels, L., & Kaplan, F. (2002). Aibos first words: The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32.

Wellens, P., Loetzsch, M., & Steels, L. (2008). Flexible word meaning in embodied agents. *Connection Science*, *20*(2–3), 173–191.

Xu, L. (1998). RBF nets, mixture experts, and Bayesian Ying-Yang learning. *Neurocomputing*, *19*, 223–257.

Yu, C., & Ballard, D. H. (2003). *A computational model of embodied language learning* (Tech. Rep. No. TR791). University of Rochester.