Grounded Language Acquisition: A Minimal Commitment Approach

Sushobhan NAYAK[1,2] *Amitabha MUKERJEE*[2]
(1) Stanford University, CA, USA
(2) Indian Institute of Technology Kanpur, India
nayaks@stanford.edu, amit@iitk.ac.in

ABSTRACT
We take up the challenge of learning a grounded model of language when our agent has a body of machine learning algorithms and no prior knowledge of either the physical domain or language, in the sense of "least commitment". Based on a 2D video and co-occurring raw text, we demonstrate how this cognitively inspired model segments the world to obtain a meaning space, and combines words into hierarchical patterns for a linguistic pattern space. By associating these two spaces under temporal co-occurrence constraints, we demonstrate the acquisition of term-meaning pairs for names, actions and relations. We next map physical arguments for actions and relations to syntactical constructions resembling a cognitive grammar framework. Thus the system is able to bootstrap a rudimentary lexicon and syntax. While experiments are primarily in English, we present partial results for Hindi obtained without any change in the methods, to indicate its potential application to other languages.

KEYWORDS: Cognitive grammar, image schema, ADIOS.

Proceedings of COLING 2012: Technical Papers, pages 2059–2076,
COLING 2012, Mumbai, December 2012.

1 Language learning: The minimal-commitment approach

We investigate a minimal-commitment approach to learn both a grounded lexicon and some rudimentary grounded syntax of an unknown language. By *minimal commitment*, we wish to restrict the prior knowledge available to our learning agent to a minimal set of abilities, and almost no resources or models of the language or domain. By *grounded lexicon* we would like to learn a bipolar relation between a unit of language and a perceptual pattern, and *grounded syntax* refers to a similar mapping from syntactic patterns to relations or events in the perceptual space (Langacker, 1987). Further, given the minimal prior knowledge formalism, the set of perceptual schemas that constitute models for meaning are obtained from the visual input in an unsupervised manner. The input to the work consists of visual sequences with simple shapes, and a set of narratives of this situation generated by adult subjects. We focus on the relation of *containment* (A in B), and the event of *chase*, and show how constructions corresponding to these are learned.

There are many works dealing with the grounded learning of words (Roy and Reiter, 2005; Siskind, 1994; Steels, 2003; Regier, 1996). Our input for word learning however, is significantly more challenging, since the narrative is a set of sentences from an unconstrained narrative, and the lexical and syntactic choices as well as the referential intentions of the speakers vary considerably. This problem is partly resolved by enabling the agent with a bottom-up model of dynamic attention, which has been shown to help computational simulations of word learning (Yu and Ballard, 2007; Mukerjee and Sarkar, 2007).

A more significant difference with earlier work is that target of the reference (a set of "concepts") is not given, but has to be discovered. A few approaches (Regier, 1996; Roy and Reiter, 2005) do discover some aspects of the semantics, but the structure is given. Thus, all approaches to grounding permit the agent to have some knowledge of the task domain in order to constrain the structures for the conceptual space; some provide a set of predicates outright (Siskind, 1994; Bergen et al., 2004; Dominey and Boucher, 2005; Caza and Knott, 2012). Let us illustrate the difficulty of making no semantic commitment with the example for containment. Without the convenience of a pre-defined predicate, the relation of being contained (the target for "in") cannot be known *a priori* but has to be first discovered as a distinct cluster in some sensory space. However, we demonstrate that such clusters emerge at least for some of the concepts of interest. In this work we simply use mean-shift clustering, but in other situations we have found that the presence of intrinsic goals can substantially improve the discrimination (e.g. for the containment task, the agent may have an intrinsic goal of inserting an object into an orifice such as a mouth).

Since the perceptual discovery operates independent of language, we assume that a set of such characterizations are already available at the start of the linguistic association process. The availability of such proto-concepts also has strong cognitive plausibility; infants are able to discriminate situations from 3 months onwards, and by 9 months, it is this ability to cluster data into groupings that lead human infants to discover the phonemic structure of their language (Mandler, 1992). One of the observations of this work is that this pre-linguistic proto-semantic discovery enables a set of categories that are specific to the domain and the goals of the agent, and eventually lead to a set of predicates that are more relevant to the situation and hence more likely to appear in linguistic discourse. Thus, this unsupervised discovery process forms the scaffolding on which the bootstrapping process works.

Attempts to learn grammatical structure range from attempts that ignore semantics altogether

to richly grounded models. The purely syntactic forms have been quite successful in inducing probabilistic grammars for tasks such as machine translation (Marino et al., 2006), and analysis of such as n-grams and path alignment have been used to determine grammars from single language corpora as well (Solan et al., 2002). In our work, we use the approach based on simple n-grams, as well as the more sophisticated (Solan et al., 2002) model, to identify the candidate syntactic structures that will be associated with the proto-semantic structures to discover constructions.

Grammar inductions that model the semantics are given a semantic structure which is matched to user narratives, obtained for instance while performing a task. The visual inputs are analyzed using a vision system into the actions identified, and these are then used to induce some aspects of grammar. This is used to learn some grammatically distinct structures (such as active or passive voice, or prepositional terms) in (Dominey and Boucher, 2005). Another body of work considers formal logical description of scenes, and induces probabilistic grammars by unambiguous sentence pairing. An impressive gain in this area has been to reduce the commitment to language knowledge, so that grammatical structures of questions in languages as different as Turkish and Japanese can be learned (Kwiatkowski et al., 2010). However, the scalability of a process that requires a large numbers of predicate specific training sets limits the scalability of the method. The objective of minimizing commitment to prior language models is essentially aimed at modeling the ability to acquire any ambient grammar.

The present work differs from all these in minimizing the dependence on prior knowledge of either the perceptual space or the language being learned. Both the perceptual structures as well as syntactic structures are obtained using unsupervised techniques, and the association performed thereafter. One important observation is that the semantics helps narrow the corpus to those sentences uttered while a specific concept is in focus, thus helping acquire structures related to it.

1.1 Capabilities of the learning agent

Start Frame	End Frame	Subject 1	Subject 2
617	635	the little square hit the big square	they're hitting each other
805	848	the big square hit the little square	and they keep hitting each other
1145	1202	the big square goes inside the box; (and) the door closes	another square went inside the big square

Table 1: Sample descriptions of events. Note the differing referential and lexical choices.

We may now define the capabilities of our minimal commitment language learning agent. We assume the agent has a) a wide range of machine learning algorithms, b) some awareness of the mental state of other agent (*Theory of Mind (Mukerjee and Sarkar, 2007)*), c) task-independent (bottom-up) dynamic perceptual attention, d) a mechanism for fixing goals (intrinsic motivation). In addition we also assume the agent has the ability to segment words from the linguistic inputs. We note that possibly such an ability may have been based on earlier exposure to the target language.

The input for our agent is a video sequence (based on Heider/Simmel (Heider and Simmel,

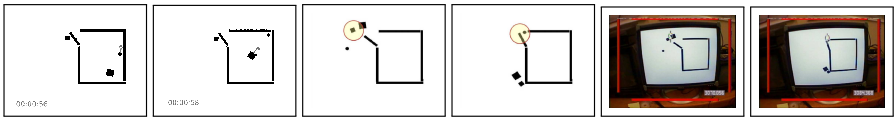

Figure 1: *Perceptual input: 2D video based on Heider/Simmel. Also showing dynamic attention model*. Three rigidly translating shapes, a big-square (**[BS]**), a small-square (**[SS]**) and a circle (**[C]**) interact playfully (velocities shown with arrows). Part of the container, a door (**[D]**) opens or closes at times. Figures 3 and 4 show the synthetic gaze computation, and Figs 5 & 6, actual gaze data for a viewer, showing reasonable correlation to predicted gaze.

1944), Fig. 1)[1]. The English linguistic database consists of 40 commentaries, collected from subjects who were simply asked to "describe the video", while they differ in certain respects, the speakers were not constrained in any form in terms of lexical choice, focus, or other aspects of their narratives. A group of 13 was collected as part of (Hard and Tversky, 2003) were asked to comment on a fine-grained vs coarse-grained temporal segmentation of the video. The other 27 narratives were collected by us. All narrators were students in the 20-25 age-group. The English narratives constitute a corpus of 4200 words (700+ sentences), and exhibit a wide range of linguistic variation both in focus (perspective) and on lexical and construction choice (see Table 1). Both the visual and the linguistic input are unlabelled.

It is possible that better results may be possible with a stronger social emphasis than was maintained in this work (e.g. joint attention vs individual attention). However, the requirement that an actual human be present makes it more difficult to scale up. As it exists, the system can possibly be tuned for a number of domains and languages.

2 Language acquisition as association

The problem of language acquisition with minimal commitment has two phases. In the pre-associative phase, the problem is to identify the semantic and syntactic primitives independently.

- *Perceptual structure discovery*. Given a perceptual space W, discover the set of structures Γ in this space, possibly focusing on high-frequency situations, or those that are relevant to its goals. Some patterns $\gamma \in \Gamma$ are simple (e.g. a shape that moves rigidly), versus others that are encode relations over space or time. Some of the simpler patterns participate in the more complex interactions.
- *Linguistic structure discovery*. Given a linguistic space L, the system is exposed to a set of sentences, each a sequence of words (w). It attempts to identify sequential patterns Λ (possibly hierarchical) that, would enable a more compact description of the input.

Strictly speaking the strong independence of the perceptual and linguistic spaces is required only for a small set of initial phrase-meaning mappings; subsequently, mappings that are known to be present in the situation can substantially constrain other possibilities, leading to great efficiencies in acquisition (Yu, 2008; Bloom, 2000) (the vocabulary spurt). In this work, we are focused on the very first, bootstrapping steps, so we maintain a strict independence. We note however, a small aspect of this independence. We observe that both linguistic and conceptual

[1]This video was developed and some of the narratives collected by (Hard and Tversky, 2003). We are grateful to Barbara Tversky for permission to use this data in this different enterprise.

structures are hierarchically organized. Thus, the a relation *A contains B* is defined over objects A and B, which are themselves structures on their own right. Thus, it is reasonable that A and B would be brought to awareness (reified) before the relation of containment. Hence, if the names of two participating objects are known, then we do use this knowledge to constrain phrases that may describe a relation between them.

We note that each sentence in the input is uttered over a particular temporal window, and includes a set of linguistic patterns, $\lambda_1, ...\lambda_m$. If the perceptual patterns observed during this utterance interval are $\gamma_1, ...\gamma_n$, then one may expect some of these λ to be mapped to some of the γ. Over many narratives describing similar situations, we are able to access a large set of such association candidates; and the association process merely posits some of the strongest co-occurrences in this large set. This association task may now be defined: Map fragments of language to the patterns of the perceptual data. Discover mappings from the meaning space to the syntax space, i.e $\gamma \in \Gamma \mapsto \lambda \in \Lambda$, where γ, λ are the perceptual and linguistic patterns that co-occur during a sentence-utterance duration.

For each pattern γ or λ, the system should be able to determine if a given perceptual or linguistic situation is an instance of the particular pattern or not. Since the initial discovery of these patterns are based on unsupervised methods like clustering, binary discriminations can be performed by bayesian techniques on the two distributions; and new instances assigned to a suitable class, or a new cluster may be initiated. In general, this means that each distribution may change significantly as more experience accumulates, but in this work, we shall not be expanding our repertoire of concepts so we shall not encounter this situation.

We note that what we are learning on the linguistic side is far from what is normally understood by syntax. We are learning merely a map from the sentential space S to a pattern space Γ, which is chosen so as to induce the largest structures that do not cause contradictions. As we shall see, these structures will use hierarchies which look rather like syntactic categories, but are quite different from traditional parts of speech or other treatments. However, we note that there can be many grammars that explain a given set of sentences, and the grammar that describes this particularly input may differ substantially from human-crafted grammars.

We observe that the grounding - i.e. the availability of a mapping to a space outside the set of logical tokens - is a very crucial part of the process by which the initially learned mappings are expanded on in further usage. Without it we would not be able to constrain the linguistic parts that are relevant when a known mapping arises(Langacker, 1987; Bergen et al., 2004).

In the first pre-associative stage, we need to define the perceptual structures. In the video (Fig. 1), the referent objects – a big square, a small square and a circle (from now on referred to as **[BS]**, **[SS]** and **[C]** respectively) – are a set of rigid translating pixels, and are easily segmented. We note here that while we claim a low degree of domain dependence, had the video been 3D or the agents been humans, it would have imposed considerable difficulties. Thus, to an extent, the perceptual analysis is limited to this kind of domain, but nonetheless, there is nothing specific to the scene. The segmented objects constitute the first level of the perceptual abstraction, and despite the variation in the names for these objects in the narrative, associating these is much simpler compared to the multi-object relations or actions.

3 Association measures

The learning objective is to create mappings between the meaning space Γ and the sentential space Λ. Suppose $\Gamma = \cup_i \gamma_i$ and $\Lambda = \cup_i \lambda_i$ be the random variables denoting the said spaces.

Each γ_i denotes a particular hypothesis for a realization of the meaning space. For example, in a verb meaning space created through clustering of motion features, γ_i would denote a cluster. Similarly, for a reference resolution task containing objects o_i, $\gamma_i = o_i$. The λ_i's similarly define partitions of the sentential space, which, based on context, might denote monograms, bigrams or other syntactical structures derived from the purely linguistic input.

We follow a statistical approach to map the two domains to each other (Fazly et al., 2010)[2] instead of an inductive approach(Siskind, 1996), since, not only ours in an unconstrained commentary, but knowing even the position of the referring expressions would be unhelpful because of noise introduced in any single isolated evidence. Though they do not use unconstrained linguistic input, there is much research on cross-situational associations between words and their meanings(Frank, 2010; Roy and Pentland, 2000; Smith and Yu, 2008; Yu et al., 2005; Yu and Ballard, 2007). We employ two association measures to correlate the meaning and sentential spaces. The *relative association*, a Bayesian metric, is defined as

$$P(\gamma_j|\lambda_i) = \frac{P(\lambda_i|\gamma_j)P(\gamma_j)}{P(\lambda_i)} \propto \frac{P(\lambda_i|\gamma_j)}{P(\lambda_i)} = A^{rel}_{ij}.$$

While working well for frequent linguistic elements, the metric is prone to give erroneous results for rare occurrences. For instance, it gives a maximum value of 1 to the correlation between a word w, which has been uttered only once in the whole discourse, and the meaning it has co-occurred with. We, consequently, also employ an information theoretic measure, the *mutual association*, defined as

$$A^{mut}_{ij} = P(\lambda_i, \gamma_j) \log \frac{P(\lambda_i, \gamma_j)}{P(\lambda_i)P(\gamma_j)}$$

because it's the contribution of each (λ_i, γ_j) pair in the mutual information of Γ and Λ

$$I(\Gamma, \Lambda) = \sum_i \sum_j P(\lambda_i, \gamma_j) \log \frac{P(\lambda_i, \gamma_j)}{P(\lambda_i)P(\gamma_j)}$$

It might also be noted that while A^{rel}_{ij} was inadequate for low frequency words, A^{mut}_{ij} gives unusually high scores for highly frequent words like *the* which have only syntactic relevance, due to a high $P(\lambda_i, \gamma_j)$, thus supporting the use of both measures for the investigation.

The goal of this work, however, is not to discover which association measure works best for word learning. The idea we are trying to support is that all categories – nouns, verbs, relational prepositions etc. – can be mapped to their corresponding meaning space through simple associations. Many association measures have been proposed in literature, which work differently in different situations. It seems cognitively implausible that an infant uses only one kind of association to discover the linguistic element with the highest correlation with the meaning and learns that as the label. Neither are many association measures too different in their output, except for a few artefacts. We instead propose that instead of looking for a *perfect* association metric to discover a single label, a more plausible approach would be to discover a *label set*, which would be refined through further syntactical, perceptual or social evidence. Since Bayesian and information-theoretic associations are well-known in NLP, we take the above

[2]However, while (Fazly et al., 2010) use an artificial meaning space for word-meaning correlation, with the meaning space created from sentential input only and essentially treated as a given, we form the meaning space from perceptual input.

[BS]			[SS]			[C]		
word(s)	A_{ij}^{rel}	A_{ij}^{mut}	word(s)	A_{ij}^{rel}	A_{ij}^{mut}	word(s)	A_{ij}^{rel}	A_{ij}^{mut}
square	0.70	1.41	little	0.66	0.79	circle	0.79	2.11
big	0.89	1.11	small	0.72	0.63	square	0.41	1.54
box	0.69	0.78	square	0.46	1.12	little	0.68	1.22
the big	0.87	0.71	small square	0.93	0.53	the little	0.71	0.81
big square	0.94	0.75	little square	0.89	0.46	little circle	0.91	0.60
large square	0.86	0.15	the little	0.70	0.54	the big	0.48	0.61

Table 2: *Noun label learning*: Word associations for the referent objects in attentional focus.

two as representative. As we shall see, both work in most situations. Note that we make no assumptions on the category of words or distinguish them as nouns/verbs. However, different machine learning algorithms are required for learning perceptual objects, events, and spatial relations; thus these are distinguished in the semantic space.

3.1 Noun reference resolution

Noun learning is well known to be easier than verbs(Fleischman and Roy, 2005). In our case, the rigid shapes are easily segmented and tracked, and the mapping of object labels to their visual representation is achieved with the help of a bottom-up dynamic saliency model. Many computational approaches have been proposed for grounded word learning(Iida et al., 2011; Prasov and Chai, 2008), though instead of working on unconstrained utterances, the referring phrases are a given. (Fang et al., 2009) uses static referents in game-like contexts, as opposed to our dynamic referents. We observe that unlike many of these approaches, our learning agent is exposed to complex discourse in which phrases are embedded, and not by isolated one-word labels.

We use visual attention to constrain the region of visual interest and identify the constituents participating in an utterance. In fact, past works like (Prasov and Chai, 2008; Iida et al., 2011) have used gaze cues from speakers to conduct reference resolution. In our case, however, since the learner is presented with only the visual stream and is not in the presence of the speaker, attention is mediated by visual saliency alone, and not by cues received from the speaker's gaze. Therefore, to simulate gaze-based visual attention, we follow the assumption of Perceptual Theory of Mind(Mukerjee and Sarkar, 2007), that the salient features that our cognitive agent discovers through image processing, would also be salient for the speaker involved in the associated commentaries, letting us correlate the visual and linguistic elements coherently. (Mukerjee and Sarkar, 2007) uses a bottom-up visual attention model to predict the gaze, the results of which are shown in Figure 1. This works as our eye-gaze model for the perceptual input. The salient agents being attended to constitute the meaning space Γ, with γ_i = object feature set. For example, the hypothesis denoting **[SS]** might represent γ_{SS} = [color:`black`, size: `25 × 25 px.`, shape: `square`, orientation: `NIL`].[3] Notice that this schema varies according to the number of features the agent is capable of deriving. The object in visual salience is then correlated with utterances that have temporal overlap with the object in focus. Since we do not assume any syntactic information at this point, every linguistic element

[3]Notice that shape: `square` is a high level concept. At present, the model can only determine through image processing techniques that length = breadth and **[BS]** and **[SS]** are of similar shape.

C1 (Come-Close)		C2 (Move-Away)		C3 (Chase)		C4 (Chase)	
move	0.033	chase	0.066	chase	0.479	chase	0.371
toward	0.028	away	0.025	try	0.115	try	0.106
corner	0.023	move	0.022	start	0.093	run	0.050

Table 3: Associating language labels to action clusters from the unsupervised algorithm

is a possible label for the objects. The association between mono- and bi-grams with the objects, for both the association measures, are shown in Table 2[4].

3.2 Verb acquisition

Once the objects are discovered, the next step is to derive perceptual relations for their interaction with each other and the surrounding. We next turn to modeling the mutual interaction between moving agents (for our input, [BS], [SS], [C]). We assume that our artificial agent is capable of employing basic unsupervised machine learning on image data, particularly the ability to segment a picture/frame, generate spatial features and cluster them into separate classes, to achieve the above mentioned goals.

Attempts to learn verbs have involved neurally inspired models of contact actions (x-schemas, (Bailey, 1997)); or a set of actions and their visual parses (Siskind, 1994; Dominey, 2005). As mentioned earlier, our perceptual schemata are discovered based on unsupervised clustering in the perception space. In this case, our perceptual extraction process depends on a specific feature that is not discovered, but these are fairly general and involve the product and difference of relative position and velocity. In earlier work (Satish and Mukerjee, 2008), we show that temporal data mining on the sequence of these feature vectors, based on Merge Neural Gas (Strickert and Hammer, 2005), yields four action clusters, two of which correspond to [come-closer] and [move-away], and two correspond to [chase].

These clusters constitute the hypothesis space for verb acquisition. These are next related to the linguistic input. For this, those sentences, which overlap temporally with the period when the action clusters are active, are taken into account, using an approach similar to (Roy and Reiter, 2005). At this point, it is assumed that the learner knows the nouns (discovered in Section 3.1), which are not considered as labels for verbs. Extremely frequent words (e.g. *the*, *an* etc.) are also dropped from consideration for mapping to actions. The strongest associations for the action clusters are shown in Table 3, with Clusters 3 and 4 (**[chase]**) having a strong association with the word *chase*.

3.3 Perceptual schema for containment

In spatial reasoning, there have been several attempts at defining spatial relations involving continuum measures defined over different geometric features on object pairs. Regier(Regier, 1996), a seminal work in preposition grounding, uses angle measures and a connectionist

[4]Notice that one word association alone might not provide sufficient information since some objects might be referred to through phrases, the validity of which is considered in Section 4.1. Also, to compare both the measures side by side, the mutual association has been scaled appropriately. The most frequent monogram *the* has been ignored in these results, which has the highest A_{ij}^{m} in all the four cases. It would later be eliminated from the probable label set anyway as it provides no information due to its occurrence in all the four label sets.

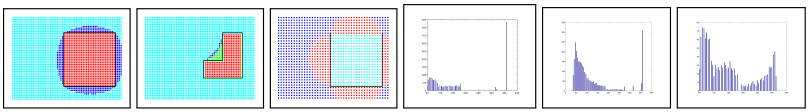

Figure 2: *Clustering through spatial features*: Figs 1-3 represent Visual Angle feature clusters. The inside of all the containers has been clearly identified as a separate cluster only in the latter case. Figs 4-6 are the visual angle histograms.

network to correlate videos and prepositions. The work, however, is limited in the sense that Regier uses videos annotated with single words like IN, OUT, THROUGH etc. while we hope to learn these schemas by clustering the untagged video. Also, because his videos are tagged with prepositions, he never has to work to *discover* the preposition; we have to discover these units from the unconstrained unparsed narrative. (Mukerjee and Sarkar, 2007) use the same dataset as ours, but use a measure based on visual proximity - the *Stolen Voronoi Area* - to cluster space using Kohonen SOMs. We initially tried these two approaches and found that in unsupervised clustering tasks (k-means and mean-shift), these earlier models do not work well for distinguishing the inside and outside of irregular (L- or U-shaped) containers. In a supervised scenario they show good results training with sophisticated neural-nets over multiple epochs, but our goal is to try not to use supervision data.

Another feature implicated in place learning in animals is *visual angle*- the angle subtended by a landmark on the retinal image. We attempted to improve on the previous features by using a single feature – the total angle subtended by a landmark at the object position. With this measure, we find that when the resulting feature space is clustered, one of the clusters works quite well for identifying the IN-schema. Computing this feature involves computing the angle that the landmark, **[box]**, would subtend at each point in the space; the result is measured and clustered using *Mean-Shift*(Fukunaga and Hostetler, 1975), so as to get non-parametric natural clusters. We can see in Fig 2 that one cluster completely covers what may be thought as the inside of **[box]**, whereas the the outside is graded between a number of clusters. If we accept this as a characterization for an image schema for containment, then the distribution of visual angle in this cluster(say the *IN-cluster* or $\mathscr{C}_{in}$) will serve to represent this relation. To test whether this model really represents the *category* of containment relations, we generalize and evaluate it over a number of other shapes. The results of clustering on two novel shapes is shown in Fig 2. We find that regions with varied levels of 'IN-ness' have been separately grouped, validating our choice of features. While for closed convex shapes the measure has a clear demarcation of 'inside'(360° angle), as is evident from the angle histograms in the figure, it gives a more graded assessment for open figures as well, such as the open-top square.

The clusters are the hypothesis space for spatial schemas. The correlation of the prominent IN-cluster($\mathscr{C}_{in}$) with words is shown in Table 4. The sentence space contained all the utterances that occurred when any of the objects in attention was inside the IN-cluster. Of interest to us is also the *change in state*, so that sentences overlapping with the object in attention moving in/out of the IN-cluster are also considered separately(Table 4, results to the right). Words *in/inside/into* are prominent, as are *into,enter* and *out,leave* for transitions in/out of IN-cluster.

4 Linguistic construct acquisition: Rudiments of syntax

At this stage, the agent is aware of some word-meaning mappings; a cognitively plausible incremental approach would suggest that the first glimmers of sentential constructions would

IN	A_{ij}^{rel}	A_{ij}^{mut}	INTO	A_{ij}^{rel}	A_{ij}^{mut}	OUT OF	A_{ij}^{rel}	A_{ij}^{mut}
inside	0.79	11.78	into	0.82	6.98	out	0.65	5.71
into	0.90	9.43	inside	0.53	1.03	leaves	1.00	4.16
in	0.61	4.16	enters	1.00	4.85	exits	1.00	3.46

Table 4: Associating language labels to the prominent IN/containment cluster

form around these recognised words. In fact, children's initial syntactic representations may be centered around individual verbs/relational items, instead of fully abstract grammars(Tomasello, 2003). There has been much work in describing such structures(Mintz, 2003; Saffran et al., 1996). The consensus seems to be that concrete n-grams or patterns ('constructional islands') like 'in the box' emerge first, with these being generalised to abstract syntactic construction like 'in the X' through distributional information in the linguistic input. We, consequently, started our discovery of syntactic structure by analyzing bi- and tri-gram correlations for containment. While the prominent bi-grams were *inside the*, *into the* and *in the*, the tri-grams that emerged were *inside the box*, *in the box* and *into the box*.

Despite some glimmers though, the n-gram approach is not very illuminating regarding the construction encoding for containment. We specifically avoid standard parsers since we are unsupervisedly discovering syntactical structures, not trained word-class labels. We, consequently, follow a richer model of syntactic structure in the spirit of cognitive grammar(Langacker, 1987), ADIOS(Solan et al., 2002), which integrates statistical and classical (generative, rule-based) approaches to syntax. It constructs syntactic representations of a sample of language from unlabeled corpus data unsupervisedly. It first creates a Representational Data Structure(RDS) by morphologically segmenting the input sentences and creating directed edges between vertices corresponding to transitions in the corpus. It then repeatedly scans and modifies the RDS to detect significant patterns through a Pattern Acquisition(PA) algorithm. A pattern tagged as significant is added as a new vertex to the RDS graph, replacing the constituents and edges it subsumes, the process being repeated and bootstrapped. Two representative patterns found through a run of ADIOS through the whole English commentary set are presented below, which provide the first indications of grouping of words in to syntactical classes:

$$1.\ \left[\begin{array}{c} the \rightarrow \left[\begin{array}{c} big \\ large \end{array}\right] \rightarrow square \\ the \rightarrow square \end{array}\right] \rightarrow \left[\begin{array}{c} scares \\ approaches \\ chases \end{array}\right] \rightarrow \left[the \rightarrow \left[\begin{array}{c} small \\ little \end{array}\right]\right]$$

$$2.\ \left[\begin{array}{c} the \rightarrow \left[\begin{array}{c} ball \\ box \\ door \\ square \end{array}\right] \\ circle \\ it \end{array}\right] \rightarrow \left[\begin{array}{c} moved \\ moves \\ runs \end{array}\right]$$

4.1 Syntactic classes refine object labels

From Pattern 2 notice that *circle, square, box, door,it,he* (say Group 1) belong to an equivalent class. Similarly, combination of words like *the big square, the little square* etc. are syntactically similar to Group 1 (Pattern 1 & 2). *open, move* etc. , on the other hand, are syntactically completely different from the aforementioned words. Also notice that *big,little,small* by themselves are not equivalent to Group 1; but as part of a bigger phrase, like *the big square*, they are equivalent to Group 1. Similar is the story with *the*. The noun label-sets from Table 2 are now curtailed to Group 1 words only, due to their high individual and combined associativity. So, while bi- and tri-grams like *small square*, *the square* are retained due to their syntactical equivalence to monograms, utterances like *door closes, circle moves* are treated as argument structures and discarded as labels. Thus, while many of these structures may offend linguists who feel they are overfitted to this particular input, it cannot be denied that it is a working model for characterizing path level agglomerations in the input.

However, in the mean while, we would like to emphasize that the process of deriving syntactic information (this Section) and mutual association of linguistic and perceptual elements (Sec 3) are not mutually exclusive or ordered processes. Even though we have described syntactic information discovery after we have motivated perceptual to linguistic element mappings, we do not assume that they are ordered that way. In fact, being independent events and mutually informative, they might as well run parallelly.

4.2 Verb and relational argument structure

Other important *containment and chase-specific* patterns, however, are hardly discovered due to the presence of myriads of different structures leading to the diffused nature of the dataset. We, consequently, follow the cognitively plausible incremental approach. We isolate those parts of the corpus that co-occur with containment/chase situations in the perceptual input, with a view that unsupervised analysis of this sub-corpus discovers regularities that are more *specific* than can be achieved under the same computational constraints for a broader corpus. The knowledge of linguistic elements *in/into/inside* and *chase* from word-label association helps constrain the corpus to a focused sub-set of 107 sentences for IN(each containing one of those words) and 36 sentences for CHASE, which facilitates discovery of some prominent structures:

1. [Group 1 word/phrase] → [CHASE (*chases/is chasing*)] → [Group 1 word/phrase]
2. [CHASE (*chased*)] → [by] → [the] → [little]
3. [CHASE (*chases*)] → [little] → [Group 1 word/phrase]

1. [Group 1 word/phrase] → [IN] → [the]
2. [Group 1 word/phrase] → [verb] → [IN] → [the]
3. [Group 1 word/phrase] → [verb] → [IN]
4. [Group 1 word/phrase] → [verb] → [other linguistic elements] → [IN]
5. [IN]→[the]→[Group 1 word/phrase]

	CHASE			IN				
Pattern No.	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**5**
Frequency	17	3	2	10	26	7	10	36
'ground-truth argument' frequency	20/34	1/3	1/2	10	18	5	9	27
'ground-truth argument' % age	59	33	50	100	69	71	90	75

Table 5: *Correlating perceptual and linguistic argument structure-CHASE & IN*

In fact, this practice of constraining hypothesis space to facilitate quick learning is supported in literature(Siskind, 1996), albeit in different contexts. Before proceeding though, we would like to emphasize that the claim is not that these are not the only patterns that are possible:[5] Our focus here is to investigate how amongst the set of plausible patterns, some are preferably acquired due to evidence and bias and strong correlation to perceptual domain. The above patterns each include at least one object term(Group 1 term) and one relation/verb term, following the hypothesis that out of patterns emerging from purely statistical data, the patterns that have a previously learned label, might be favorably acquired.

While these structures have been derived from purely linguistic input, their *grounding*(bootstrapping in perception) is possible only if they show remarkable correlation with the perceptual argument structure, which they indeed do, as can be ascertained from Table 5. In the table, 'ground-truth argument' means the perceptual and linguistic agents are in conformation. Conflict cases involve both when the linguistic agent is different from the perceptual agent (e.g. in video [CHASER] is **[BS]**, but in the utterance, [CHASER] is *circle*) and when the linguistic agent is unfamiliar (e.g. in video [CHASER] is **[BS]**, but in the utterance, [CHASER] is *big block – block* is syntactically equivalent to *square* so that the structure is valid, but the agent has not *yet* associated it with any perceptual objects from past evidence). Since Structure 1 has two referents, the total number of referents for 17 sentences is 34. While raw correlation is greater than 50%, if we discount the sentences with unfamiliar linguistic agents, there is 100% correlation between linguistic and perceptual schemas, thereby making the linguistic argument structure concrete.

All the IN patterns show more than ∼70% correlation between the two domains, leading to the grounding of respective structures. Note that the attention based model for learning nouns cannot learn the container/box, which is never dynamically salient. Thus, it's label is unknown. However it is prominent in these containment sentences, and discounting the frequent word *the* in trigrams such as "*{ inside/in/into} the box*", we may associate *box* with **[box]**, treating it as a label for the container. It logically follows since **[box]** is a physical object, and based on the past experience of the agent, should be assigned a linguistic element that syntactically confirms to concepts of other physical objects[6] like **[BS], [SS]** etc. From the above patterns, the syntactical equivalence of *box* is supported by its grouping with *door, square* etc.(see Pattern IN4). This grouping into equivalent classes is the first evidence of word category acquisition. The primary mapping of *box* to **[box]** is further strengthened from Table 5, where *box* has been taken as the 'ground-truth argument' for **[box]** and with this assumption, more than 75% of the IN2 pattern sentences agree in both perceptual and linguistic domain, thereby facilitating the

[5]In fact, from a statistical viewpoint, with change in length of input and variation in ADIOS parameters, myriads of patterns can emerge.

[6]For the present set-up, all the moving objects and the container can be derived through image segmentation, thereby being similar 'physical' objects.

[BS]			[SS]			[C]			[IN]		
word(s)	A_{ij}^{rel}	A_{ij}^{m}	word(s)	A_{ij}^{rel}	A_{ij}^{m}	word(s)	A_{ij}^{rel}	A_{ij}^{m}	word(s)	A_{ij}^{rel}	A_{ij}^{m}
बक्सा baksA/box	.77	.37	बक्सा baksA/box	.62	.44	गोला golA/ball	.83	.54	अन्दर andar/in	.80	1.30
बडा(badA/ big) बक्सा	.85	.18	छोटा(chota/ small) बक्सा	.90	.25	बक्से के(ke/–)	.63	.27	बाहर (bA-har/out)	.78	.73

Table 6: *Noun label learning*: Word associations for the referent objects in attentional focus.

acceptance of the assumption. In fact, of the 9 mismatched referents, only 3 are wrong (*in the door/corner/place*, while the rest are due to synonymous references (*in the room*, *in the square*).

One possible consequence of grounded syntax is a facilitation of acquisition of minor labels, synonyms and anaphoras, which are overwhelmed by other salient labels in a simple association task. While we have not entered into a serious investigation of this, some rudimentary results are apparent nonetheless. *ball* and *room* occur 5 and 6 times respectively in positions A and B in sentential domain in 'A in/inside/into B' and map exclusively to $x =$ **[C]** and $y =$ **[box]** for xINy in perception, so that they are treated as synonyms for *circle* and *box* (in this scenario). *block* has been used for both **[BS]**(75%) and **[SS]**(25%), creating an equivalence with *square*. Similarly, *it* maps to **[BS]**, **[SS]** and **[C]** in 56, 25 and 19 percent of its occurrences, thus showing the first evidence of anaphora acquisition, though we leave a detailed investigation to a future work. How these acquired structures can further be extended to assimilate metaphors and how they are developmentally salient for the concept of containment, has been investigated in much detail elsewhere (Nayak and Mukerjee, 2012).

Potential application to other languages To investigate the potential extension of this approach to other languages, we obtained results for Hindi obtained without any change in the methods. The Hindi database consisted of 10 commentaries from first-language speakers, with more than 200 sentences and 2000 words, describing the same video (e.g. तो लगता है कि यहां एक बडा बक्सा है जिसमें एक चौकोर है [to lagtA hE ki yahAN ek badA baksA hE jismeN ek chokour hE / `It seems that there is a big box here in which there is a square present.`]). Hindi is a much more richly inflected language than English, with abundance of gender and number agreements. Even verbs have several modal affixes in addition to tense and, several postpositional markers for case, which sometimes mark for source and destination as well. The constructions derived from ADIOS, therefore, are diffuse and minimal, owing to the small dataset(200 sentences, compared to 700 in English). So, while a detailed investigation of syntax grounding is far-fetched at present, the emerging patterns show consistencies with their English patterns nonetheless (*comes/runs out of box*):

$$\begin{bmatrix} \text{डब्बे(dabbA/box)} \\ \text{बक्से(bakse/box)} \end{bmatrix} \rightarrow \begin{matrix} \text{के} \\ \text{(ke/–)} \end{matrix} \rightarrow \begin{bmatrix} & \text{बाहर(bAhar/out)} & \\ \text{बाहर} & \begin{bmatrix} \text{आ(aa/come)} \\ \text{भाग(bhAg/run)} \end{bmatrix} & \text{जाता} \\ \text{(bAhar/out)} & & \text{(jAtA/goes)} \end{bmatrix}$$

Furthermore, similar results to English are found for word-to-meaning mappings (See Table 6). Also, we have पीछा(pichA/chase) as the dominant label for the verb clusters, with $(A_{ij}^{rel}, A_{ij}^{mut})$ of $(0.78, 19.79)$ and $(0.64, 13.7)$ respectively for Clusters 3 and 4 from Sec 3.2.

Conclusion and perspectives

The work described is able to learn a limited set of lexical items and grammatical constructions for a small domain. If it would be possible for the system scale up, one may suggest that such a system might be able to learn an increasing number of concepts from a larger number of domains, as it happens with children. Indeed, the holy grail of computational linguistics would be to create such semantically rich models of language, and alternatives like this approach are at the very least worth investigating further. But how difficult would it be to scale up this computational approach to new domains and new concepts?

The reason why minimal commitment is attractive is of course, precisely because it makes it easier to extend the approach to other physical domains or other languages. Situations where other concepts have saliency would make it more likely that associations with linguistic expressions mapping them would be learned. Also, with the capacity for simulation, it is no longer necessary to have direct grounding for learning everything. In a sentence where a novel term or concept is introduced, the meaning of the term, or the concept itself, may be understood by simulating what is known from the other parts of the expression. Indeed this is how humans learn the vast majority of our immense vocabularies (Bloom, 2000).

What the initial term-meaning pairing provides is an index into the space of meanings. The next time a similar expression is encountered, the new semantics are compared with the previous mental model so it can be extended. Thus if our system here, which knows "in the box", now encounters "in the basket" and "in the room", it would gradually be able to generalize the argument of "in the" to the concept of container (and would eventually reject "in the banana", say). Further, as we have demonstrated elsewhere (Nayak and Mukerjee, 2012), once the system encounters increasingly figurative expressions such as "in the team", "in the school", "in the spotlight", "in the doghouse", it would be able to extend this using the analogy mechanism inherent in the sensorimotor schemas used for grounding. We emphasize again the need for sensorimotor grounding without which such sense extensions are clearly not possible.

This work presents a view that takes seriously and implements computationally, the ideas in Cognitive Grammar (CG) (Langacker, 1987), as opposed to traditional grammars. This view is part of the models in (Regier, 1996; Bergen et al., 2004; Chang and Maia, 2001), but ours is the first to propose a model that also learns the semantics, thus freeing it up to learn new structures in new domains. What is really being learned in this process is what is called an image schema in CG and for larger structures, a more elaborate schematization. With the grounded model, as we encounter increasingly complex perceptual schema that map to longer phrases, it should also be possible to discover the processes of composition in CG.

At this point, such claims may seem too remote, but they are not completely implausible and given their potential for changing the way NLP works today, we would argue at least for the widespread development of corpora with image streams along with multiple raw text narratives in many situations and many languages. We make a humble start in this direction by making our image and text corpora in several languages available on the web. Agents that accumulate the learning from several domains may prove (or disprove, or suggest new directions) in the enterprise. The main claim we are making is that this is a novel approach to language induction, and while its potential is far from clear, at least it is worthy of further investigation.

References

Bailey, D. (1997). *A Computational Model of Embodiment in the Acquisition of Action Verbs*. PhD thesis, UC Berkeley, Dept EECS.

Bergen, B., Chang, N., and Narayan, S. (2004). Simulated action in an embodied construction grammar. In *Proc. of the 26th Annual Meeting of the Cognitive Science Society*.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

Caza, G. and Knott, A. (2012). Pragmatic bootstrapping: a neural network model of vocabulary acquisition. *Language Learning and Development*, 8(2):113–135.

Chang, N. and Maia, T. (2001). Grounded learning of grammatical constructions. In *AAAI Spring Symp. On Learning Grounded Representations*.

Dominey, P. (2005). Emergence of grammatical constructions: Evidence from simulation and grounded agent experiments. *Connection Science*, 17(3-4):289–306.

Dominey, P. F. and Boucher, J.-D. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, 167(1-2):31–61.

Fang, R., Chai, J., and Ferreira, F. (2009). Between linguistic attention and gaze fixations in multimodal conversational interfaces. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pages 143–150. ACM.

Fazly, A., Alishahi, A., and Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Scince*, 34(6):1017–1063.

Fleischman, M. and Roy, D. (2005). Why verbs are harder to learn than nouns: Initial insights from a computational model of intention recognition in situated word learning. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

Frank, M. (2010). *Early word learning through communicative inference*. PhD thesis, Brain and Cognitive Sciences, Massachusetts Institute of Technology.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IT, IEEE Transactions on*, 21(1):32 – 40.

Hard, B. and Tversky, B. (2003). Segmenting ambiguous events. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57:243–259.

Iida, R., Yasuhara, M., and Tokunaga, T. (2011). Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proc. of IJCNLP*, pages 84–92.

Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., and Steedman, M. (2010). Inducing probabilistic ccg grammars from logical form with higher-order unification. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1223–1233.

Langacker, R. (1987). *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press.

Mandler, J. M. (1992). How to Build a Baby .2. Conceptual Primitives. *Psychological Review*, 99(4):587–604+.

Marino, J., Banchs, R., Crego, J., de Gispert, A., Lambert, P., Fonollosa, J., and Costa-jussà, M. (2006). N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91 – 117.

Mukerjee, A. and Sarkar, M. (2007). Grounded perceptual schemas: Developmental acquisition of spatial concepts. In *Spatial Cognition V Reasoning, Action, Interaction*, volume 4387, pages 210–228. Springer Berlin / Heidelberg.

Nayak, S. and Mukerjee, A. (2012). Learning containment metaphors. In *Proc. of CogSci*.

Prasov, Z. and Chai, J. (2008). What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 20–29, New York, NY, USA. ACM.

Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Bradford Books.

Roy, D. and Pentland, A. (2000). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26:113–146.

Roy, D. and Reiter, E. (2005). Connecting language to the world. *Artificial Intelligence: Special Issue on Connecting Language to the World*, 167:1–12.

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Satish, G. and Mukerjee, A. (2008). Acquiring linguistic argument structure from multimodal input using attentive focus. In *ICDL 2008*, pages 43 –48.

Siskind, J. (1994). Grounding language in perception. *AI Review*, 8:371–391.

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

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

Solan, Z., Ruppin, E., Horn, D., and Edelman, S. (2002). Automatic acquisition and efficient representation of syntactic structures. In *Proc. of NIPS*.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308–312.

Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, 64:39–71.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language learning and Development*, 4(1):32–62.

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

Yu, C., Ballard, D. H., and Aslin, R. N. (2005). The Role of Embodied Intention in Early Lexical Acquisition. *Cognitive Science*, 29(6):961–1005.