1. Introduction

Framing the problem of language acquisition, which has traditionally been the niche of psychology and linguistics(Gold, 1967; Pinker, 1989), as a machine learning problem has helped formalize it in the following manner(Kaplan, Oudeyer, & Bergen, 2008):

- The learner's objective is to map space A to space B.
- A set of paired points, $x \in A \mapsto y \in B$, are the training set on which the learning is performed.

Consequently, the learning agent has to first discover the map from the sentential space S to binary space of syntactic correctness B, thereby assimilating the structure of the language. Subsequently, the agent has to discover the bidirectional mapping between S and the meaning space M. This framework has begotten much research in the computational approaches to linguistic concept acquisition.

These computational approaches, however, have been partial towards the discovery of the first kind of mapping, i.e. the syntactical relevance of sentences. While the generative stance(Chomsky, 1975; Gold, 1967; Chomsky, 1957) has regarded syntax learning as the central acquisition problem, the statistical stance(Elman, 2005; Solan, Horn, Ruppin, & Edelman, 2005) has favored the accumulation of statistical and probabilistic linguistic information over hypothesis testing for grammatical rules. Both the stances are critical of each other. The generative stance believes in the Poverty of Stimulus(POS) hypothesis, which champions the idea of a Universal Grammar that is innate to a learner, since the input available to children is inadequate for language learning(Chomsky, 1975; Gold, 1967). The statistical stance considers use of language, the actual production and comprehension of utterances, and some(Elman, 2005) argue that children are privy to more information than previously assumed, repudiating POS. A celebrated argument favoring POS and criticizing statistical methods is due to Chomsky(Chomsky, 1957). Chomsky proposed two sentences, `Colorless green ideas sleep furiously` and `Furiously sleep ideas green colorless`, and argued that while neither of the sentences has a meaning, only the former can be viewed as being "grammatical", so that grammaticality judgments hold independent of meaning and syntax is autonomous from semantics. Further, since either sentence may not have ever been used in English discourse, this judgment of grammaticality is not based on a statistical models, thereby refuting claims of the rising trend of statistical grammars. This criticism of statistical methods, however, is ill-conceived, because it can be regarded as an attack on not the statistical method *per se*, but *n-gram* models

of grammar acquisition, which assign probabilities to co-occurring linguistic elements. Under that assumption, both the above sentences are low or zero probability word sequences. When we move above the co-occurrence of words model, and consider instead the syntactic relation at word-class level, the phenomenon can be supported by statistical methods. While the first sentence is a traditional Noun phrase-Verb phrase sentence construction, the second sentence is VP - NP and the NP in the second sentence has the ordering of Noun-Adjective, instead of the usual Adjective-Noun formation. In fact, (Pereira, 2000) has shown that the later sentence is 200,000 times less probable than the former, conforming to Chomsky's observation, from a statistical point of view. The point we are trying to make is, statistical analysis might seem helpless only under narrow constraints, but it has tremendous power once the domain of the problem is correctly recognised. Furthermore, as (Elman, 2005) has claimed, while POS is based on inadequacy of evidence from linguistic input only, there might be further evidence available to the learner, evidence, we believe, that comes from embodied and social cognition.

While the previous two stances were primarily focused on the syntactic mapping, the embodied stance tries to address the second mapping problem, that from sentential space to the meaning space. While the previous two stances essentially ignore the meaning and embodied aspect, "yet, the fact that language and concepts are acquired through the use of a physical body and the fact that language is an interactive process taking place between individuals situated in social and physical environments could not help but constrain language learning"(Kaplan et al., 2008). Embodied investigations have primarily remained theoretical in nature(Lakoff & Johnson, 1980; Langacker, 1987; Bergen, Chang, & Narayan, 2004), with minimal computational investigation, primarily because of the complications arising out of the expansion of the scope of investigation beyond isolated and idealized linguistic knowledge. Such approaches, however, have been made, where models investigate acquisition of spatial or action linguistic concepts through evidence from perceptual domain(Regier, 1996; Bailey, 1997) and through physical robots(Steels & Kaplan, 2001; Dominey, 2005). The embodied approach not only captures the word-meaning mapping, but it also proposes ways to handle both the sentence to syntax and meaning mapping through a common framework(Langacker, 1987; Bergen et al., 2004). *Cognitive Grammar*, due to Langacker(Langacker, 1987), takes the radical position that grammar reduces to the structuring and symbolization of conceptual content and thus has no autonomous existence at all. According to this, language necessarily comprises semantic and

phonological structures and symbol links between the two. The central claim of Cognitive Grammar is that nothing else is needed.

It's evident that different stances propose sometimes varying interpretations of the language acquisition task and focus on different aspects of the problem. In view of the foregoing discussion, it seems that a meaningful computational model that can address the question of acquisition and can adhere to combine both the mapping of sentence and syntax and sentence and meaning, would be one that appeals to both the statistical and embodied aspects of language. We are of the vies that the generative stance can be subsumed by a combination of both the statistical and embodied stance; while word/word-class statistics provide information about favorable sentential structures, the embodied meanings provide the missing evidence from extra-linguistic perceptual and sensual input. Let's revisit the seemingly meaningless sentence `Colorless green ideas sleep furiously`. We argued previously that the seemingly correct grammar can be explained from statistical description alone, without invoking Universal Grammar, if the probabilistic methods are applied at the sentence structure level or word-class level, instead of relying solely on word-co-occurrence. We go on to show that, an embodied approach also can make the sentence meaningful. The seminal work by (Lakoff & Johnson, 1980) brought forth the idea that metaphors, instead of being a 'device of the poetic imagination and the rhetoric flourish', an integral part of our action and thought processes. Abstract linguistic elements are mapped to embodied concepts of Object, Substance and Containers, concepts that emerge directly for an early learner through physical experience. "We experience ourselves as entities, separate from the rest of the world - as containers with an inside and outside. ... We experience ourselves as being made of substances-... (Lakoff & Johnson, 1980)." From that perspective, 'colorless green ideas' would be mapped to objects, activating a mapping of IDEAS ARE OBJECTS metaphor, which is an established mapping in English (Ch. 10,(Lakoff & Johnson, 1980)). Thus, we see that a combination of both stances might be capable of addressing problems inherent in different independent approaches. In the description that follows, we propose such a system and show its capability in acquisition of linguistic concepts, albeit in a primitive way. To demonstrate the power of grounded acquisition, we further go on to show that the combined approach can further help us discover and acquire anaphoric and metaphoric (especially containment metaphors) linguistic elements. Consequently, a sentence like `Colorless green ideas` ***`that`*** `sleep furiously` ***`in mind`***, which encompasses linguistic aspects of syntax, grammar, and anaphora(*that*) and figurative aspects of metaphors(IDEAS ARE PEOPLE/OBJECTS, MIND IS A CONTAINER), makes sense.

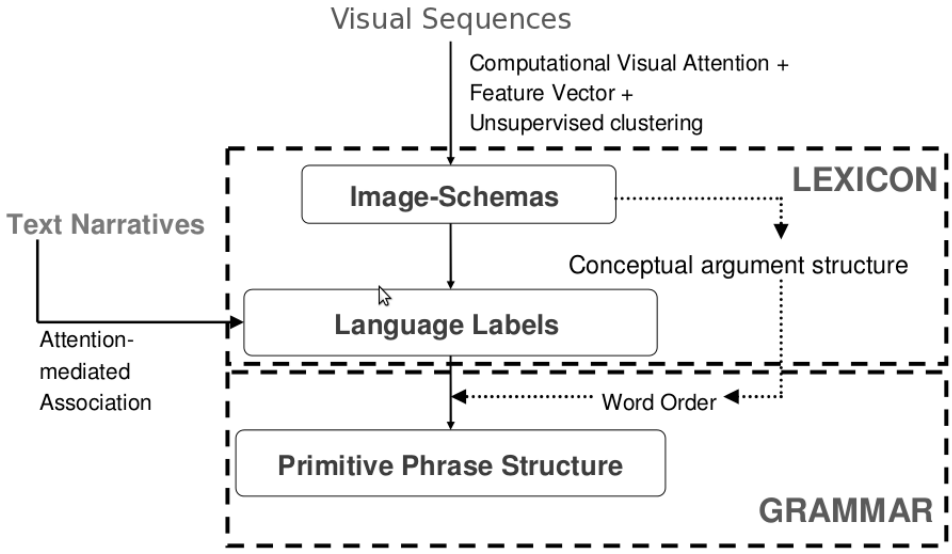


Figure 1: *Acquiring a grounded language.* Linguistic structure discovery relies on similarity in sensorimotor space. Our approach discovers image-schemas from perceptual input unaided by language, and then associates these unsupervised notions with units from text narratives. It then uses these semantic associations to map metaphoric and anaphoric usage.

1.1 Towards a Cognitive Model

Statistically mapping structure to meaning based on cross-situational inference was first described in (Siskind, 1996). The present work is posited along such lines. While Siskind assumes the existence of certain constraints that limit the hypothesis space of meanings, we derive such constraints from perceptual input. This embodied approach, though intellectually appealing and elegant, is hard to scale up due to two reasons. First, it is enormously difficult to posit new semantic structures for the vast array of linguistic concepts. Secondly, even given the semantics, it is difficult to relate it to language without some knowledge of syntactic structure and how it relates to this semantics. In this work, we attempt to present a grounded model where the rich semantic models underlying linguistic transfers need not be carefully engineered, but may be learned based on simple sensori-motor inputs (Roy, Hsiao, & Mavridis, 2004).

In our approach, we consider perceptual inputs in terms of image sequences and discover a set of sensorimotor object, action and relation clusters, and then associate these with words from text (Fig 1). The action and relation clusters are image schemas and a cluster such as IN would also identify objects that are likely to be containers. Next, the syntactic structures are discovered (in a

very primitive form) from this set of narratives, and structures such as `the big square chases the` {`circle, little square`} or `the` {`circle, square`} {`is, are`} `in the` {`box`} emerge. This constitutes the Grammar aspect in Fig 1. We then further show how such acquired information can be used for metaphor and anaphor handling. Essentially, through multi-modal input consisting of a video and co-occurring commentaries, we try to simulate the experience of an early learner. Literature abounds in studies that follow this multi-modal approach to language learning(Roy & Reiter, 2005; Fang, Chai, & Ferreira, 2009; Steels, 2003; Siskind, 1996; Dominey & Boucher, 2005). While some of them aim to identify the referents in an interaction discourse(Fang et al., 2009; Steels, 2003; Siskind, 1996), others often use prior knowledge for visual parsing of actions(Dominey & Boucher, 2005; Siskind, 1996). Since our process is completely unsupervised, it would be possible to scale it up by applying it to novel situations. Also, while our approach is similar in spirit to that of (Regier, 1996) and (Bailey, 1997), the above works are limited in the sense that they try to map single word sets to meaning sets, while we also address which words out of an unconstrained set would be likely to be assigned to the embodied meanings. To demonstrate the generalness of applying such grounded linguistic knowledge to general text, we search for structures similar to what we have discovered in the Brown corpus, and identify several instances of metaphorical usage related to the "in the X" construction. Thus, this entire process, starting from a multimodal input, is able to discover perceptual spatio-temporal pattern, map them to linguistic units, relate these to a primitive notion of syntax, and exploit this conceptual basis to acquire metaphorical and anaphoric mappings in a natural way that emulates language learning in an early learner. In the following section, we detail the unsupervised acquisition of pre-linguistic action and relational schemas.

2. Discovering Prelinguistic Image Schemata

2.1 The Cognitive Agent

Linguistic concepts are cognitively characterized in terms of *image schemas*, which are schematized recurring patterns from the embodied domains of force, motion, and space(Langacker, 1987; Lakoff & Johnson, 1980). The precise structure of an image schema remains quite unclear, with different authors using differing characterizations. In this work, we take an image schema to consist of two related structures. First is the list of arguments which participate in the associated relation or activity. The other is a characterization of the situation in terms of a function defined over some

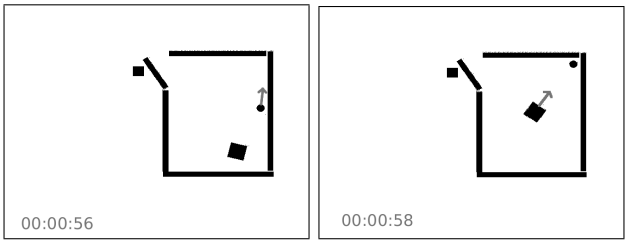


Figure 2: *Multimodal input: 2D video "Chase"*: Three easily segmented shapes, a big-square ([**BS**]), a small-square ([**SS**]) and a circle ([**C**]) interact playfully (velocities shown with arrows). Part of the landmark, a door ([**D**]) opens or closes at times.

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: *Description of the Events by Subjects*. Differing statements by two subjects in the video.

feature space, so that situations satisfying this function may be considered as instances of the image schema. For our primitive agent, we therefore consider a perceptual database with multi-modal input capabilities, viz. a 2-D video and co-occurring commentaries(Heider & Simmel, 1944)(Figure 2), which simulate the visual and auditory senses. In the video, the referent objects – a big square, a small square and a circle (from now on referred to as [**BS**], [**SS**] and [**C**] respectively) – are moving around, interacting with each other, and are easily segmented, as opposed to static referents in game-like contexts used in other multimodal co-reference analysis (Fang et al., 2009). The door ([**D**]) of the landmark, a box([**box**]) opens or closes at times to contain or let go one of the above three. The linguistic database consists of a co-occurring narrative with 40 descriptions of the video. In the 13 from the original Stanford corpus[1], subjects were asked to discriminate actions in a fine and coarse manner. The subsequent 27 collected by us, also from student subjects in the 20-25 age-group, were completely unconstrained. Thus, these narratives exhibit a wide range of linguistic variation both in focus (perspective) and on lexical and construction choice(see Table 1).

Besides the obvious relational advantage of visual and linguistic nature, we chose this set up primarily for two reasons. Firstly, the model we use, tries to learn semantic and syntactic features

1. This video was developed and the narratives collected by Bridgitte Martin Hard and Barbara Tversky of the Space, Time, and Action Research group at Stanford University (Martin & Tversky, 2003)

of the linguistic input. And cross-situational learning, which we are going to follow, is ambiguous in general. Children apply cross-situational inference for word learning, constraining the possible meanings of words, given their context of use(Siskind, 1996). However, the visual context is ambiguous in the sense that the salient agents participating in the acts might be of complex geometrical shapes, leading to confusion in separation. As an example, learning `dog` while observing a **dog** might also induce **fur**, **tail** etc.(Yuan & Xu, 2011). (Siskind, 1996) deals with it using some compositionality assumptions and constraining hypotheses with partial knowledge. Our use of simple easily segmented figures as input allows us to keep the possible number of referents in control so that no need for assuming any such cognitive faculty on the part of the primitive agent arises; at the same time, the use of three different objects in the video keeps the reference resolution and word learning tasks challenging. Secondly, using unconstrained dialogue simulates child language acquisition more faithfully. Some(Locke, 1975; Bruner, 1983) argue that learning happens with single-word utterances in a context where the meanings of those words are made clear by ostention. However, as (Siskind, 1996) points out, while only 1913 (5.6%) out of the 34,438 utterances in the NINA corpus in CHILDES database(MacWhinney & Snow, 1985) are single-word utterances, fewer than 30% of the parental utterances (trying to teach their children words) consists of isolated words, supporting our choice for the unconstrained narrative as opposed to constrained narratives used in some other works(Regier, 1996; Bailey, 1997; Prasov & Chai, 2008).

With this set-up, the objects involved in the video are easily discovered. Though the objects deform a bit while rotating, and also occasionally overlap, it is relatively straightforward to segment them. Once the cognitive agent is capable of distinguishing participating objects, it would naturally take notice of the action they are involved in. In the present video, there are mainly two distinct kinds of actions involved, to wit, actions in which the objects interact with each other (for example, chasing, hitting each other etc.) and actions in which they interact with the background/landmark, thereby bringing in spatial aspects(especially containment) to focus. We will presently show that there are plausible groundings for motion and spatial schemas, which can then be correlated to the words in the commentary for action label/verb and spatial preposition learning, subsequently helping us in extracting syntacto-semantic structures from uttered sentences, thereby facilitating anaphora recognition and resolution and metaphor acquisition.

2.2 Discovering Perceptual Action Structure

Mapping action labels/verbs to their respective motion schemas has been detailed in previous works(Mukerjee, Vaghela, & Shreeniwas, 2004; Mukerjee, Neema, & Nayak, 2011). In these works, using monadic and dyadic motion features from the video, mappings for words like `chase`, `spin`, `come closer`, `move away` were created. Here, we briefly review our previous work(Mukerjee et al., 2011) on verb acquisition. We focus on the problem of spatial preposition acquisition in the following section(Section 3.3).

We detail how an unsupervised process may acquire action structures from simple videos by clustering frequently observed sequences of motions. Two-agent spatial interactions, which correspond to verbs with two arguments, are considered. The model uses bottom-up dynamic attention (Figure 7, Section 3.1) to identify the objects that are related by attention switches. The system considers pairs of objects attended to within a short timespan, and computes two inner-product features

1. The relative-velocity and relative position

$$pos{\cdot}velDiff : (\vec{x}_B - \vec{x}_A) \cdot (\vec{v}_B - \vec{v}_A)$$

2. The relative pose and the sum of the velocities

$$pos{\cdot}velSum : (\vec{x}_B - \vec{x}_A) \cdot (\vec{v}_B + \vec{v}_A)$$

The temporal histories of these feature vectors are then clustered using the temporal mining algorithm Merge Neural Gas (Strickert & Hammer, 2005). Four action clusters are discovered, two of which correspond to **[come-closer]** and **[move-away]**, and two correspond to **[chase]**(Figure 3). Chase has two clusters because it is asymmetric, and the primary attention may be on the chaser (Cluster 3) or on the chased (Cluster 4). By computing the feature vectors with the referents switched, the system can by itself determine this alternation.

2.3 Perceptual Schema for Containment

In the previous section, we encountered linguistic elements dealing with the mutual interaction of moving objects in the video. In this section, we focus instead on spatial aspects of the perceptual

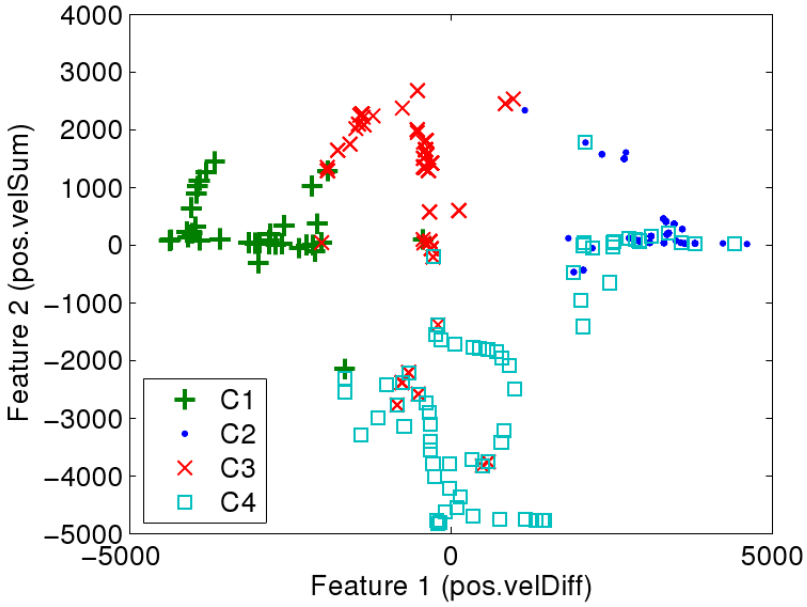


Figure 3: *Feature Vectors of the Four Clusters* : CC: C_1, MA: C_2, Chase(focus is on **[chaser]**): C_3, Chase(focus is on **[chased]**): C_4; The clusters reflect the spatio-temporal proximity of the vectors. From (Mukerjee et al., 2011)

input that arise due to interaction of the objects with the background/landmark. The particular aspect that is prominent in the present set-up is the idea of containment. The objects repeatedly go inside and leave the landmark which is in the shape of a closed square/box.

Spatial relationship concepts are often fuzzy and imprecise, consequently demanding a qualitative representation or model of space. In the spatial domain, there have been several attempts at defining spatial relations involving continuum measures defined over different geometric features on object pairs. Some of the previous works like (Mukerjee & Sarkar, 2007) use a measure based on visual proximity - the *Stolen Voronoi Area*[2] to cluster space using Kohonen Self-Organizing Maps(SOMs)(G. Edwards & Gold, 1996). Regier(Regier, 1996), another seminal work in preposition grounding, uses angle measures (esp. from the horizontal and vertical). The work, however, is limited in the sense that Regier uses videos annotated with single words like IN, OUT, THROUGH etc. , while we are looking to learn these mappings from unrestricted dialogue. We, in this work,

2. A set of partitioning lines delimiting the areas closest to an object in a group of objects is the area voronoi diagram, a partitioning of the image space based on proximity to objects, each cell in which is the set of points that are most proximal to the enclosed site. If a new object is now introduced, its measure of spatial position in terms of existing objects can be found in terms of the extent to which it reduces the areas of the existing voronoi cells - a concept which Edwards et al [11] have called Stolen Voronoi Area.

make use of both these types of features(stolen Voronoi area and (Regier, 1996)-inspired distance and angle features), and add one of our own, to wit, the *visual angle measure*, to derive containment schemas and associate them with the commentary. In the process, we learn two things. For regular containers like the square box in the video, even a subset of these used features are capable of segmenting the space into the inside and outside of the container. Consequently, even if we assume a limited capability on the part of an early learner, say its capability of only recognizing angle features or for that matter only area features, spatial schemas would be one of the prominent and distinguishing features for the agent. Secondly, for complex shapes like open boxes or L- or U-shaped boxes, our proposed feature, i.e. the visual angle measure, outperforms the previous ones as far as correct segmentation into inside and outside of the container is concerned.

In our video, the trajectors(the squares and the circle) and the landmark(the big box/ container) are easily separated, thanks to the motion features described before(Mukerjee et al., 2004). The center of mass(CoM) of the landmark, i.e. the big box, is obtained by taking the centroid of all the stationary black pixels (thereby not assuming any knowledge of a closed entity). Similarly, the simplicity of the figure allows us to detect the edges and corners of the landmark with standard image processing techniques without any assumption on the interior/exterior of the landmark, which is crucial as we are trying to *learn* containment itself. Now given that the corners, edges and CoM have been be seamlessly derived without compromising the learning integrity, we can derive, for an agent like the square or the circle, the subtended angles to the vertical from the 5 points (4 corners and the CoM of the box) and the 4 signed perpendicular distances from the 4 edges of the landmark. With stolen Voronoi area, each frame of the video can thus be reduced to a point in a 10-D space with respect to each of the three trajectors (big square, small square and circle).

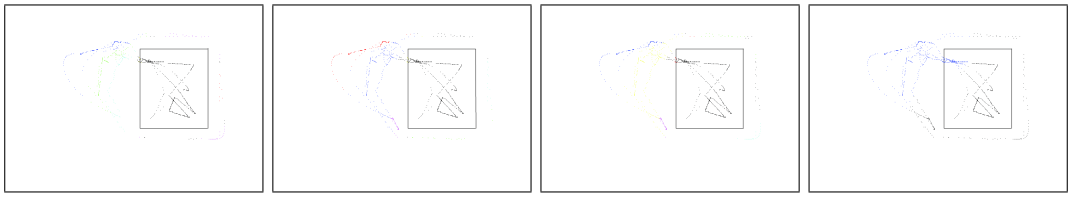
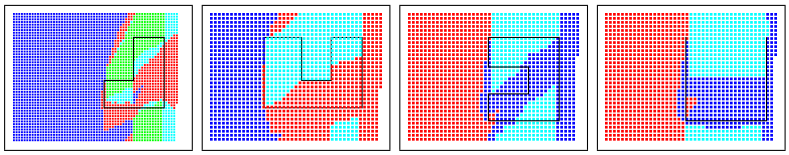
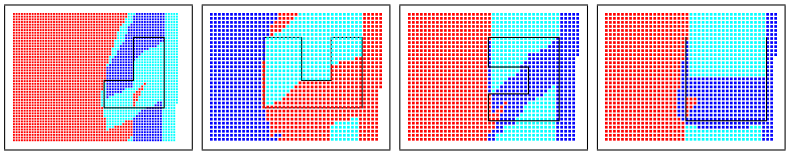


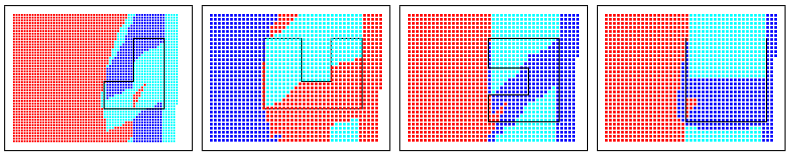
Figure 4: *Clustered trajectory of BS*: The figures respectively represent clustering results for [area+angles], [area+distances], [all 10] (they create 4 clusters each) and [angles+distances](2 clusters created). The inside of **[BOX]** is dominantly and exclusively clustered into a single dominant black cluster in the first 3 instances (showing a clear idea of containment), while the 4th, without the Stolen Voronoi area, is unable to do so, with some major part of the surrounding also being included in this cluster.



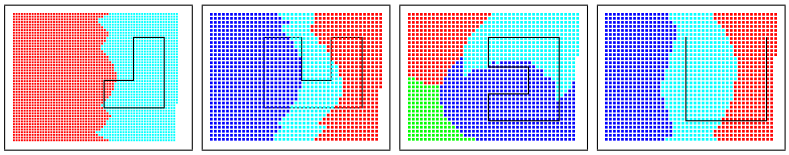
(a) Features used: area and angles



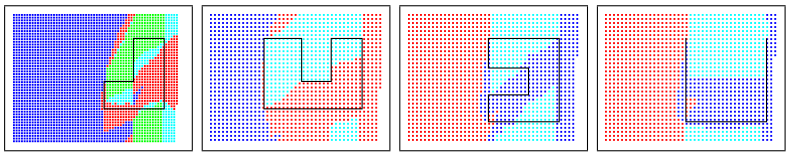
(b) Features used: area and algebraic distances



(c) Features used: all 10 – area, distance and angle



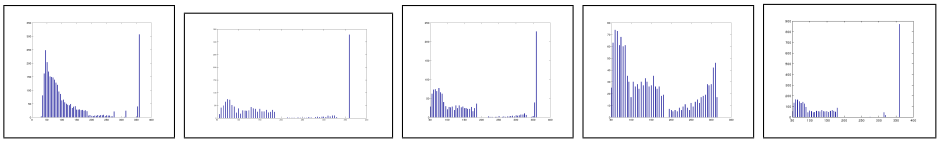
(d) Features used: angles and distances



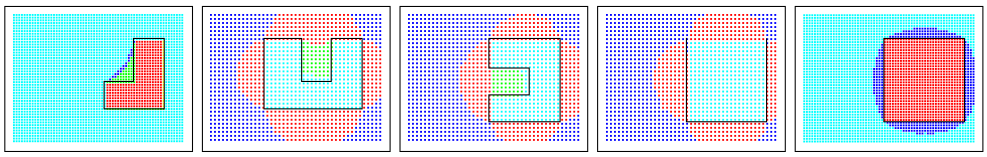
(e) Features used: stolen Voronoi area only

Figure 5: *Clustered Space*: The figures respectively represent clustering results for [area+angles], [area+distances], [all 10], [angles+distances] and stolen Voronoi area only for 4 differently shaped containers.

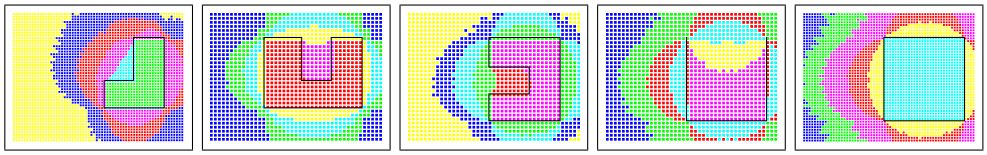
We extract these features for the Big Square[**BS**]. We use the *Mean-Shift algorithm* (Fukunaga & Hostetler, 1975), a non-parametric feature space analysis technique (non-parametricity allows us to divide the feature space into arbitrary number of clusters, creating an imperative hypothesis



(a) Angle histograms for corresponding shapes



(b) Mean Shift clustering with visual angle for the shapes



(c) KMeans-6 clustering with visual angle for the shapes

Figure 6: *Clustered Space and Histograms*: Histograms and clustering results when visual angle measure is used. Notice the clear distinction between inside and outside of the containers

space for unsupervised containment learning) to cluster the feature space. For completeness and to gauge the importance of the features, we try combinations of angle, distance and area to discover the clusters. The results are shown in Fig 4. Notice that the various combination of features give impressive results. The part of the trajectory of [**BS**] that is inside the box is almost exclusively clustered into a single cluster, giving strong evidence for a containment schema. Also note that, stolen Voronoi area features are the most important of these features, as in the fourth figure, where this feature is not used, [**IN**] and [**OUT**] clusters sometimes overlap. Nonetheless, the apparent uniformity in containment clusters across different feature spaces supports our first proposition, that a subset of these used features are capable of segmenting the space into the inside and outside of the container, so that even an agent with limited spatial information can segment space into contained and unconstrained regions. Therefore, we might assume that spatial containment is one of the earliest pre-linguistic concepts learned by a cognitive agent, which, later on, helps acquire abstract concepts of anaphora and metaphors, as we will presently show. In fact, this assumption of containment being

learned early is not far-fetched; all of a child's experiences, starting from the mother's womb to the crib, inculcate in it a sense of boundary(Lakoff & Johnson, 1980). Because the idea of containment is recognized so early on, it becomes computationally efficient to base and understand abstract ideas through this concept. Once an early learner has a grounding of containment, it can acquire syntactic information of the sentences describing the phenomenon, and consequently can detect and resolve pronomial anaphora, as we will presently show. But before we tackle that issue, another important aspect needs looking into.

While the given features are capable of separating contained space from open space for a regular container like the square box, would they perform effectively for any arbitrary shape? So, to see if a *category* for containment has been learned, we generalize from the training set to new shapes. Psychologically, changing the shape of the training objects is a standard test for determining if it is the spatial concept that has been learned, or a function limited to the specific shapes (Casasola, Cohen, & Chiarello, 2003; Spelke & Hespos, 2002). Figure 5 show results of clustering on four such different shapes for the sub-sets of features described beforehand. In these figures, feature vectors have been calculated at every possible spatial position inside the frame. As is evident, the features do not distinguish clearly between contained and free space for irregular/complex geometrical figures like L-shaped, or U-shaped containers. However, for an open-top square container, they show fairly accurate results(last shape in each subfigure of Fig 5). Points deep inside the container have been grouped into one cluster, while those closer to the opening, which have reduced 'in-ness', have been grouped into another, with outer points being still further disambiguated. It supports our previous observation that the features perform well for regular shapes. However, to deal with the general case, we propose the use of *visual angle measure*. We show that this feature alone shows promising results for all types of shapes.

For an object and a landmark, the angle subtended by the landmark at the object is taken as the subtended angle measure or the visual angle measure. It should be noted that, this subtended angle measure is fundamental in a human visual system and has a prominent role in place learning in animals is *visual angle* (Rolls et al., 1999). We get some idea about the distance of an object and it's three dimensional perspective from the angles subtended in our eyes. So, it seems a very natural and primitive feature to be taken into account. For the shapes we have considered, Figure **??** shows the angle histograms for subtended angle measures. Notice that, for a closed figure, there is a sharp peak at 360 degrees, corresponding to the points exactly inside the container, where the agent

would be completely surrounded by the container, leading to a 360 degree subtended angle. This clear demarcation prophesies a clear clustering of the space into contained and free space, which indeed does happen, as can be ascertained from Figure 6(b). See that, in all the shapes, the inside is clearly separated from the outside. For L-shaped and U-shaped figures, the outside crevices have been grouped separately, because, they also provide a sense of 'in-ness' – not the fully contained one, but analogous to the in-ness of a open container, like the open-top square. This graded in-ness is also the reason the previous features were somewhat unable to distinguish between the contained and free space for irregular shapes. This, therefore, supports our second claim that the subtended angle measure outperforms the previous ones as far as correct segmentation into inside and outside of a container is concerned. The results using the subtended angle measure for **[BOX]** have been shown in Fig 6.

At this stage, the system computes the visual angle subtended at an object position, for a landmark (the box or some other shape). The two arguments participating in this computation are the container and the trajector, though the system does not use these terms; these relationships are implicit in the feature computation. If the visual angle falls within the distribution associated with containment (the *IN-cluster*), it is accepted as an instance of this relation. Thus the system has both the arguments, and the acceptance function characterizing the image schema, This acquisition is pre-linguistic, from perceptual data alone. When a pre-discovered object, say **[BS]**, lies in IN-cluster, we have the argument structure {**[BS]** IN **[box]**} or IN(**[BS]**,**[box]**). After learning the unit IN, we can attempt to map this perceptual argument structure to linguistic syntactical structure, thereby discovering aspects of syntax.

3. Discovering Concept-Label Association

3.1 Noun-Reference Resolution

Our first objective is to learn the mapping of object labels to their visual representation with the help of a visual saliency model. Often referred to in literature as 'reference resolution', many computational approaches have proposed for grounded word learning(Iida, Yasuhara, & Tokunaga, 2011; Prasov & Chai, 2008). While some look at single objects (Roy, 2000), others(Steels, 1997) use single word utterances which can easily be correlated with perceptual precepts. We consider an immersive learning scenario where an user is exposed to multi-object scenes accompanied not

by one-word labels but by words appearing in usage contexts, requiring us to handle two issues in particular:

1. Which aspect of the scene does an utterance refer to (Quine's gavagai problem (Quine, 1960))?

2. Given that the input consists of full sentences and not single word utterances, which part of the sentence is the label for the aspect chosen in Part 1?

Using visual attention to constrain the region of visual interest and identify the constituents participating in an utterance is a natural way to go. In fact, past works like (Prasov & Chai, 2008; Iida et al., 2011) have used gaze cues from speakers to conduct reference resolution. There have been significant studies(Just & Carpenter, 1976; Griffin & Bock, 2000; Ballard & Yu, 2003) that correlate eye-gaze and speech and assert that the eye fixates on symbols that are being cognitively processed, working as a window to mind. The direction of gaze carries information about the focus of a person's attention (Just & Carpenter, 1976). 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 & Sarkar, 2007) as explained in Section 3.1.1. For identifying which part of the sentence is the most appropriate label, we compare all utterances where the object is in attentive focus, and evaluate the words most consistently being used. Thus we used no syntactic cues, *at first*. Indeed there is some agreement that language learners can acquire some nominals from language usage based on correlation alone, e.g. (Bloom, 2000) says, "Syntax is not necessary for at least some nominal learning." We describe this in Section 3.1.2.

3.1.1 Visual Attention Model

Recognizing moving objects, distinguishing shapes and predicting motion require a developed visual processing system in an agent. The paradigm that language is learnt by correlation with what's going on in the visual field (the view we subscribe to here) inherently assumes a highly developed visual system capable of image and scene processing. Therefore, since our focus is on the first steps of linguistic element detection, while we are constrained to choose a cognitive system with limited text/speech processing capabilities (to wit, in this work, the Bayesian inference), we are free to

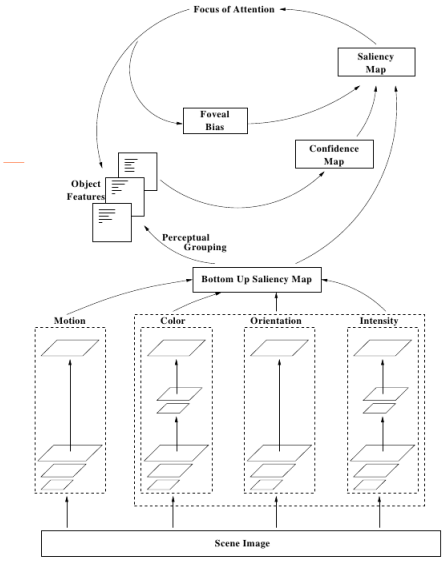
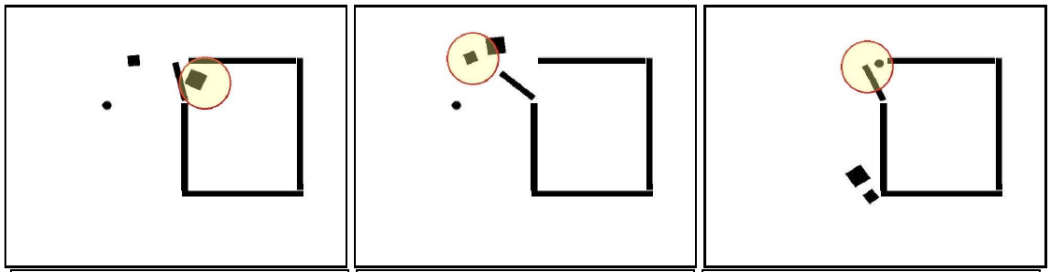


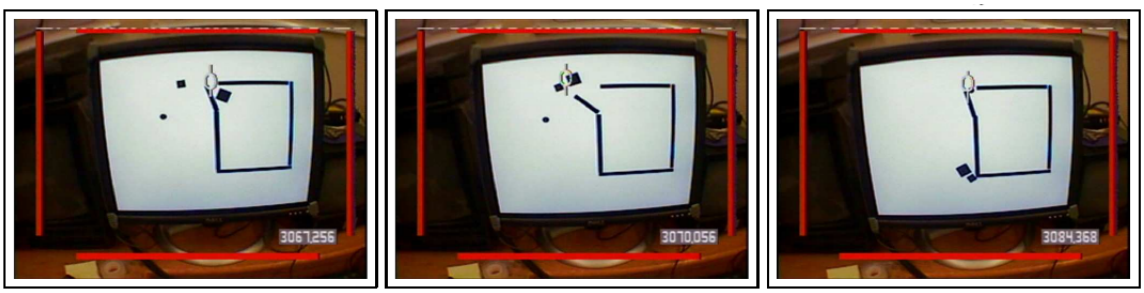
Figure 7: *Bottom-Up Dynamic Visual Attention Model.* The feature maps for static images (color, intensity and orientation) are extended with a motion saliency map (based on optical flow). In addition a confidence map records which sites have not been visited for a longer time. Winner-Take-All determines the next fixation.

assume a more complicated visual system. We use one such synthetic model of visual attention in this work.

In a previous work of our group(Sarkar, 2006), we proposed the *Perceptual Theory of Mind*. While an interested reader might want to consult that work for details, we describe here the features of the model relevant to the present work. Where as much of the Theory of Mind(Bloom, 2000) work has focused on gaze following based on cues from the speaker's eyes or her gaze direction, the Perceptual Theory of Mind makes a much weaker claim: in the absence of direct cues from a speaker, it assumes that the speaker would have attended to those parts of the scene that the learner also finds salient. An interesting consequence of this theory is 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. In our visual attention model(Mukerjee & Sarkar, 2007), color and intensity contrast maps are obtained as feature pyramids (maps at different scales), along with center-surround maps (multi-scale difference of feature maps). The center-surround feature processing is similar to the difference of gaussian convolved images (DOGs). For orientation spe- cific processing, gabor filters are used with different frequencies and at different scales to generate the orientation specific feature map.



(a) Focus of attention on the video. Red circles represent the focus of Attention



(b) The speaker's gaze tracked through the same scenes. Note the discrepancies between the views.

Figure 8: Comparison between predicted gaze and the real gaze from human experiment.

The static model, which replicates saliency map structures likely to be present in the LGN or V1 regions of the mammalian cortex, has been extended (Maji, Singh, & Mukerjee, 2006) to model dynamic scenes based on motion saliency. Motion saliency is computed from the optical flow, and a confidence map is introduced to record the uncertainty accumulating at scene locations not visited for some time. A small foveal bias is introduced to mediate in favour of proximal fixations as opposed to large saccadic motions. The saliency map is the sum of the feature maps and confidence maps, mediated by the foveal bias, and a Winner- Take-All (WTA) isolates the most conspicuous location for the next fixation. The overall architecture is shown in Figure 7".

The synthetic model of visual attention for dynamic scenes was used to predict the gaze, the results of which are shown in Figure 8(a). In Figure 8(b), we show the visual gaze from an user narrative (separate from our text data) in which the narrator is describing the scene while wearing a gaze tracker[3].While no formal comparisons have been performed, we observe that most narrators

3. These images were obtained by Vivek Singh courtesy of Dana Ballard and Brian Sullivan at the University of Rochester Cognitive Vision Laboratory.

seem to be focusing on, or shifting between, the same objects roughly in the same part of the narrative - thus providing some informal justification for our technique.

After discovering the salient agents being attended to in the visual domain, we next go on to correlate them with linguistic input for reference resolution.

3.1.2 Name Label Learning

(a) Monogram association with [**BS**]

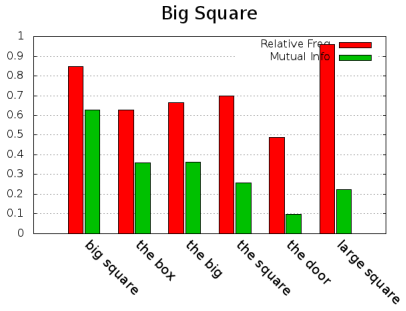
(b) Monogram association with [**SS**]

(c) Monogram association with [**C**]

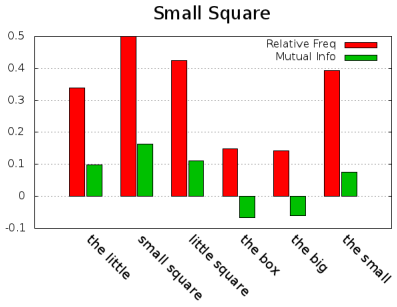
(d) Monogram association with [**D**]

Figure 9: Figure showing association of single words with the four objects in the video. Both the association measures are shown. The red bars indicate the Relative Freq measure while the green bars are for the Mutual Information Measure.

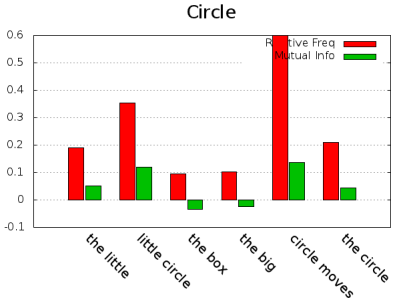
We are using unconstrained commentary to learn word label. We do not assume any syntactic information at this stage, so that every uttered word is a possible label for the agent in focus. The attended objects are therefore associated with the temporally correlated words through a Bayesian measure. We assume that the cognitive agent focuses primarily in correlating an agent with the occurring words, with the consequence that correlation with motion (verb learning) and other possible



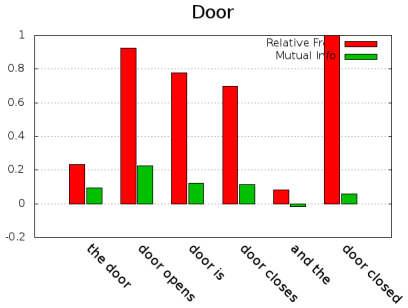
(a) Bigram association with **[BS]**



(b) Bigram association with **[SS]**



(c) Bigram association with **[C]**



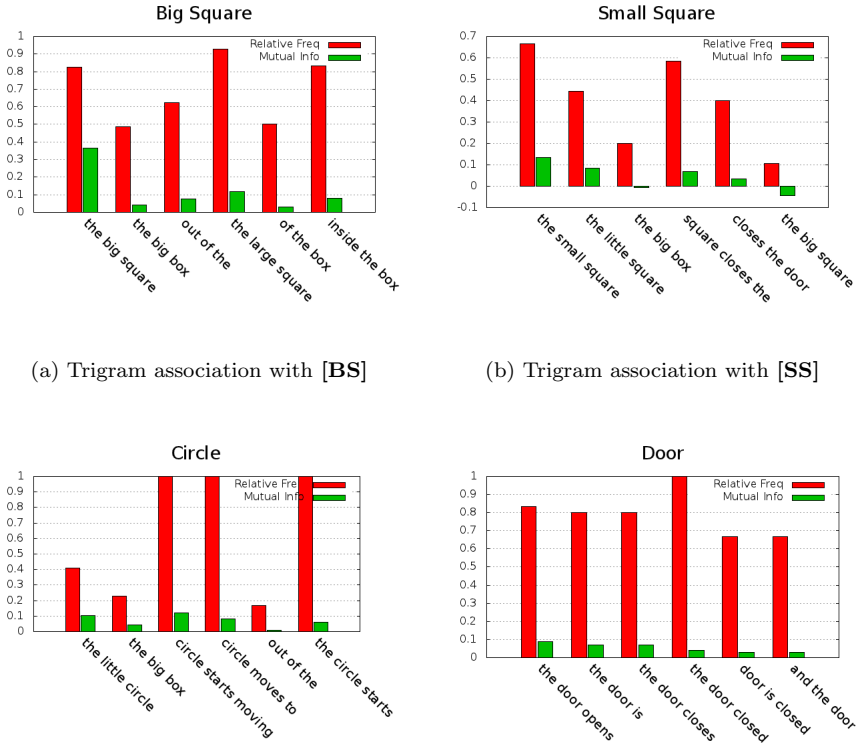
(d) Bigram association with **[D]**

Figure 10: Figure showing association of two words with the four objects in the video. Both the association measures are shown. The red bars indicate the Relative Freq measure while the green bars are for the Mutual Information Measure.

syntactic/semantic knowledge acquisition takes a back stage. Under this assumption, it is plausible that the cognitive agent tries to relate every word it comes across with the objects in the video, to find the best possible match. In that scenario, given utterances w_i, we intend to find the probability that the utterance refers to object o_j, i.e.

$$P(o_j|w_i) = \frac{P(w_i|o_j)P(o_j)}{P(w_i)} \propto \frac{P(w_i|o_j)}{P(w_i)} \propto \frac{freq\ (w_i)\ \ when\ O_j\ \ in\ focus}{freq\ (w_i)} = A^r_{ij}.$$

One might however notice that there are two issues this approach is unable to handle. Firstly, the above metric A^r_{ij}, the *relative association*, has a probable shortcoming. For rare linguistic elements, this metric is prone to give erroneous results, while working well for high frequency words. For example, it gives an association value of 1 for a word that has been uttered only once in the whole



(a) Trigram association with **[BS]**

(b) Trigram association with **[SS]**

(c) Trigram association with **[C]**

(d) Trigram association with **[D]**

Figure 11: Figure showing association of three words with the four objects in the video. Both the association measures are shown. The red bars indicate the Relative Freq measure while the green bars are for the Mutual Information Measure.

commentary. To counter this trend, we also subscribe to mutual information between objects o_j and words w_i, which eliminates the possibility of uninformative rare words being assigned a high score. The word-object association is then estimated using the product of mutual information of word w_i and object o_j with their joint probability,

$$A_{ij}^m = Pr(w_i, o_j) \log \frac{Pr(w_i, o_j)}{Pr(w_i)Pr(o_j)}$$

where A_{ij}^m is the *mutual association*. We calculate the product of joint probability and mutual information because if $W(= \cup_i w_i)$ and $O(= \cup_i o_i)$ are two random variables then their Mutual

Information $I(W, O)$ would be

$$I(W,O) = \sum_i \sum_j Pr(w_i, o_j) \log \frac{Pr(w_i, o_j)}{Pr(w_i)Pr(o_j)}$$

and $Pr(w_i, o_j) \log \frac{Pr(w_i,o_j)}{Pr(w_i)Pr(o_j)}$ would be the contribution of each word object pair.

Secondly, using one word associations alone can lead to incompletely derived information. For example, in the present video, where there are two objects of similar shape, to wit **[BS]** and **[SS]**, the objects might be (and are) referred to as `big square` or `small square` etc. so that one word association only might discover only `square`, or `big` or `small`, whereas looking for bi- or tri-grams might provide us with complete information. We, therefore, also correlate such linguistic elements from the commentaries. The results are shown in Figures 9, 10 and 11.[4]

From Figure 9(a), we see that `big` and `it` show the highest relative association, while `square` comes third, owing to the fact that `big` and `it` have been primarily used in the part of the commentary where **[BS]** is in attentive focus. So even though `it` is a low frequency word, it has a high relative association. As expected, the mutual association eliminates this anomaly, with `square` emerging as the prominent monogram associated with **[BS]**. Similarly, `little` and `small` emerge as the prominent monograms associated with **[SS]**, while `square` follows them(Fig 9(b)). The reason `square` has such low association in this case is that the occurrence of `square` in **[BS]**-specific discourse far outnumbers that in **[SS]**-specific discourse. This trend also favors the usage of more than one word for correlation. `square`, `big`, `little` etc. are only partial representations of linguistic elements for the objects. In scenarios where there are things with similar properties, like the big and the small square, more information becomes a necessity, which is not the case with the clearly distinguishable `circle` for **[C]**(Fig 9(c)) or `door` for **[D]**(Fig 9(d)). This is supported by the fact that when bigrams and trigrams are taken into consideration, correct labels emerge, with **[BS]** primarily being associated with `(the) big/large square` and **[SS]** with `(the) small/little square`.

At the same time, one might notice that, two and three gram associations for the circle and the door give away unnecessary information (`circle moves, door opens` etc.). A question that naturally emerges here is, how would a cognitive agent recognize the concept boundary; i.e. how can it demarcate linguistic units of variable length that are equivalent? Specifically, how does it recognize

4. What matters is the relative height of the bars. Therefore, to plot both the measures side by side, the mutual association has been scaled appropriately.

CLUSTER 1 (Come-Close)		CLUSTER 2 (Move-Away)		CLUSTER 3 (Chase)		CLUSTER 4 (Chase)	
ONE WORD LONG LINGUISTIC LABELS(MONOGRAMS)							
corner	0.077	away	0.069	chase	0.671	chase	0.429
move	0.055	move	0.055	other	0.185	after	0.112
attack	0.042	chase	0.049	around	0.183	out	0.033
TWO WORD LONG LINGUISTIC LABELS(BIGRAMS)							
each other	0.086	move away	0.111	chase around	0.306	chase after	0.218
move toward	0.065	go into	0.035	each other	0.227	just chase	0.060
toward each	0.065	into with	0.035	chase each	0.198	chase out	0.058
THREE WORD LONG LINGUISTIC LABELS(TRIGRAMS)							
move toward each	0.182	go into with	0.099	chase each other	0.558	just chase out	0.142
toward each other	0.182	run away out	0.051	start run away	0.132	run away out	0.047
move close together	0.114	scare in corner	0.032	begin to move	0.127	to go after	0.031

Figure 12: Figure showing the strongest association of linguistic labels and action clusters. Dominant association of [**chase**] with the word `chase` is evident.

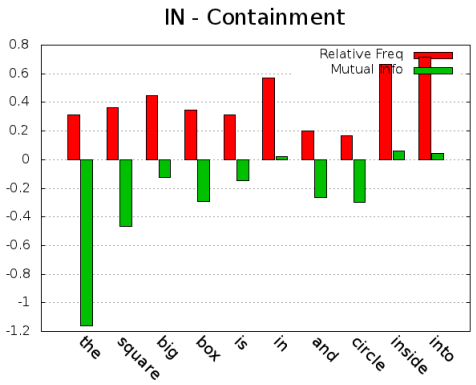
that `the big square` and `circle` are equivalent(noun/noun phrase) whereas `circle moves` isn't because `moves` is a different kind of linguistic element(verb)? We claim that this can be addressed through syntactical information, and we look into the matter in Section 4.

3.2 Discovering Action Labels

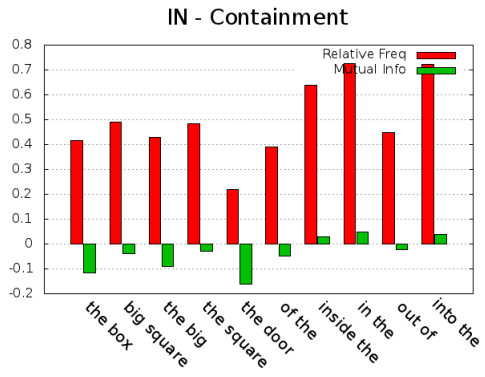
The learned models or *visual schemas* of action are acquired prior to language, and defined on the perceptual space (Section 2.2). The learned models include the agents participating in the action, which constitutes the visual arguments of the action. These are next related to the linguistic input. 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 & Reiter, 2005). At this point, we assume 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. Using 1-, 2- and 3-word sequences from the text, the strongest associations for the action clusters are shown in Figure 12, and we see that clusters 1([**come-closer**]) and 2 ([**move away**]) have strongest associations with `move toward each` and `move away`, but these are not very dominant over other competitors. On the other hand, for clusters 3 and 4 ([**chase**]), there is a strong association with the word `chase`.

We have, therefore, a correlation between spatial motion schema clusters and the linguistic elements of the commentary. However, for them to be of any help in anaphora resolution or metaphor acquisition, we would have to derive the argument structure of [**chase**], so that from linguistic input alone, the agent is capable of extracting information about the [**chaser**] and the [**chased**], an issue we deal with in Section 4.

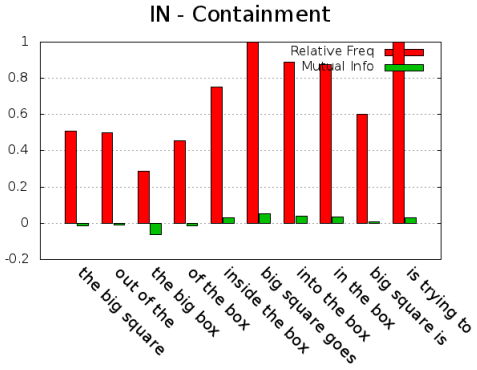
3.3 Spatial Descriptors for Containment



(a) Monogram association with the IN-cluster



(b) Bigram association with the IN-cluster



(c) Trigram association with the IN-cluster

Figure 13: Figure showing association of words with the containment cluster. Both the association measures, as used previously for noun-label learning, are shown. The red bars indicate the Relative Freq measure while the green bars are for the Mutual Information Measure.

There is strong cognitive evidence that acquisition of some spatial relations such as containment may be pre-linguistic(Lakoff & Johnson, 1980). Based on this hypothesis, we first described a plausible way to acquire perceptual schemata for containment in Section 2.3. We now detail the discovery of linguistic elements pertaining to containment through correlation of the commentary with the acquired schemata. Fig 4 shows the trajectory of [**BS**]. Of interest to us is the cluster that represents containment (the prominent black cluster in Fig 4 or the red one in Fig 6(b) last subfigure). So, all the utterances occurring during the time [**BS**] is in that part of its trajectory are considered. The same association measures, the ones used for noun-label learning, are used to correlate mono-, bi- and tri-grams from the commentary to the IN-schema. The results are shown in Fig 13. Notice that `in`, `inside` and `into` emerge as the dominant monograms.

The bi- and tri-gram correlation can have interesting consequences. We see that the prominent bi-grams are 'inside the', 'into the' and 'in the'. The tri-grams that emerge are 'inside the box', 'in the box' and 'into the box'. These associations could help us learn the label of not only the containment schema, but also the container itself. Notice that due to its static nature, the container/the box doesn't come into visual focus during the commentary, eliminating the prospect of discovering its label through the attention model. However, as a part of the landmark, it's prominent through out the video, and the idea of containment is somewhat incomplete without a label for the container. If we subscribe to Siskind's(Siskind, 1996) ***constraining hypotheses with partial knowledge*** view, " When learning word meanings, children use partial knowledge of word meanings to constrain hypotheses about the meanings of utterances that contain those words." According to this, once the cognitive agent has correlated [**IN**] to containment, it would try to associate other relevant words to co-occurring phenomena/agents. It might notice that `the` appears in almost all contexts(in fact, it's the most frequent word in the discourse and in English too), so that it's taken as a linguistic element without a concrete mapping in conceptual domain, and of only syntactic value, that might modify the semantics of a occurrence in a way yet unknown to the agent. Therefore, the agent might associate `box` with the [**BOX**], treating it as a label for the entity. However, we might notice that `big square goes` and `is trying to` also emerge as prominent trigrams. So we again come across the problem (first encountered in Section 3.1) of necessary and sufficient information for label learning. We claim that this issue can be handled with derived syntactic information from the utterances and go on to show that in the following section.

4. Deriving Argument Structure

When we are at the stage when the processes described in the previous sections have taken place, the cognitive agent has, at its disposal, the following:

- a set of labels for objects in the video that are in attentional focus
- labels describing motions/object-object interactions in the video
- a set of labels for the containment schema

For an agent with this information, the co-occurring commentaries are a set of words with recognizable labels thrown in in-between. More information can only be derived through some syntactic knowledge of the language. Specifically, we would like to discover two linguistic aspects of importance:

1. Syntactically similar linguistic elements
2. Argument structure for spatial and motion schemas

Syntactically similar or equivalent elements would help us disambiguate phrases like `the big square` from `square goes in`, so that whereas the first one is assimilated as a possible label for the objects in the video, the later is discarded as only an argument structure or linguistic pattern. Discovering argument structures of the containment preposition IN, or that of verb CHASE, on the other hand, will enable us to discover other labels for objects participating in containment, which will possibly lead to discovery of synonymy, anaphora and metaphorical mappings. For instance, the *perceptual* argument structure of IN at our disposal is IN(A,B), where A is the trajector and B is the container in the visual field. What we require is syntactic information about the *linguistic* argument structure so that the two may be mapped, i.e. we need syntactic information to convert a set of words like 'A goes into B' to INTO(A,B), even though the agent might not know what 'goes' means, so that the agent is aware of the trajectors and landmarks involved in a discourse. Once the agent learns this argument structure, not only would it be able to recreate context in a novel discourse, through derivation of probable trajectors and landmarks through information from syntactic argument structure, but would also be able to derive synonyms for objects in the present discourse by correlating objects from sentential and visual information.

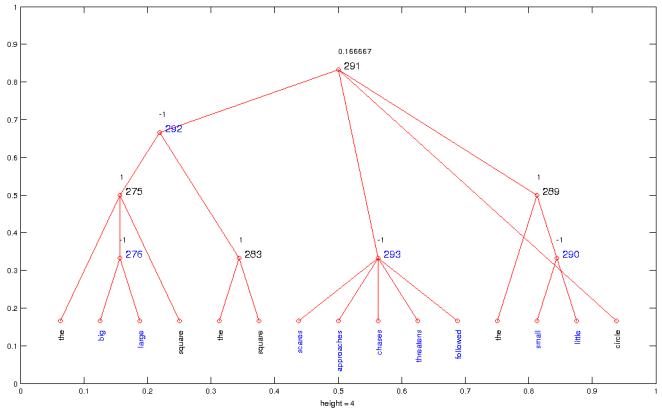


Figure 14: ADIOS output of a pattern. The elements in blue refer to the same equivalent class, whereas black labels are for patterns.

At the beginning of this discussion, we have repeatedly emphasized the assumed limited linguistic processing capabilities of the cognitive agent. We, therefore, would require an algorithm, that can derive useful syntactical information from the unconstrained narrative with minimal statistical prowess. While use of state-of-the art parsers might provide us with better results, our objective in this work is to provide a plausible way for an artificial agent/ early learner with limited capacity to acquire nuances of the language it encounters. We consequently favor ADIOS(Solan, Ruppin, Horn, & Edelman, 2002) to other parsers(Marie-Catherine de Marneffe & Manning, 2006) thanks to its claim that it finds syntactic categories from a corpus without requiring pre-judging of either the scope of the primitives or of their classification, in an unsupervised way that is cognitively simple and plausible for a child. A brief description of the algorithm follows.

4.1 A Peep into ADIOS

ADIOS(Solan et al., 2002) integrates statistical and classical (generative, rule-based) approaches to syntax. It's a scheme "that acquires"raw" syntactic information construed in a distributional sense, yet also supports the distillation of rule-like regularities out of the accrued statistical knowledge." It constructs syntactic representations of a sample of language from unlabeled corpus data and all the information needed for its operation is extracted from the corpus in an unsupervised fashion. 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. For each path p_i , the algorithm constructs a list of candidate constituents, $c_{i1}, \ldots, c_{ik}$. Each of these consists of a "prefix" (sequence of graph edges), an equivalence class of vertices, and a "suffix". Metric I', which ensures that long patterns with high mutual information between its constituents are chosen, is used to judge pattern significance :

$$I'(c_1, c_2, \ldots, c_k) = e^{-(L/k)^2} P(c_1, c_2, \ldots, c_k) \log \frac{P(c_1, c_2, \ldots, c_k)}{\prod_{j=1}^{k} P(c_j)}$$

where L is the typical context length and k is the length of the candidate pattern; the probabilities associated with a c_j are estimated from frequencies that are immediately available in the graph. 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.

The algorithm was run on the commentary corpus and one of the patterns that emerge are shown in Figure 4.1.

4.2 Discovering Syntactic Information

Some of the prominent patterns emerging from ADIOS are(for readability, output figures like Fig 4.1 were converted to the following format):

Pattern 1

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

Pattern 2

$$\begin{bmatrix} the \rightarrow \begin{bmatrix} ball \\ box \\ door \\ square \end{bmatrix} \\ circle \\ it \end{bmatrix} \rightarrow \begin{bmatrix} moved \\ moves \\ runs \end{bmatrix}$$

Pattern 3

$$the \rightarrow \begin{bmatrix} large & larger \\ little & bigger \\ other & empty \\ small & smaller \end{bmatrix} \rightarrow \begin{bmatrix} ball \\ square \\ box \end{bmatrix}$$

Our initial objective was to discover syntactically equivalent linguistic elements. From Pattern 2 notice that `circle, square, box, door,it,he` (say Group 1) belong to equivalent classes. 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`.

These groupings into equivalent classes provide us with a way to disambiguate labels for the objects. We will motivate that through an example here. Consider the single words with highest association for the four objects. Perceptually, the four objects are similar physical devices. So, it's but natural that they should all be described in similar syntactic elements. We proceed by elimination. The syntactic class `small, little, big` appears in only three of the four probable sets, while being absent from that of [**D**], so that the labels can't belong to this group. Similar arguments eliminate the verb class `is, close, open`. `the` has the lowest association in three cases, eliminating this class as a choice. The only class common to all four is Group 1. Therefore, while [**BS**]and [**SS**]

have a probable mapping to `square, box, it`, [**C**] may be assigned to `circle,box`(though the low association means `box` would never be considered a label for [**C**] even though syntactically it's a probable candidate). But [**D**] is clearly assigned to `door`.

Use of bi- and tri-grams further enriches these mappings. Of the six bigrams in Fig 10(a), `the big` isn't equivalent with Group 1 and `the door` has too low an association. The rest four, to wit `big square, large square, the square, the box`, are valid descriptions for [**BS**]. Using similar argument, `small square, little square` and `the circle, little circle` emerge as the only two probable candidates for [**SS**] and [**C**] respectively. Of the [**D**] bigrams, only `the door` is equivalent to Group 1. Similarly, trigram associations create the following mappings: ([**BS**] $\rightarrow$ { `the big square, the large square`}), ([**SS**] $\rightarrow$ { `the little square, the small square`}), ([**C**] $\rightarrow$ { `the little circle`}). Notice that erroneous phrases belonging to the correct syntactic group (e.g. `the big box` for [**C**]) fall off due to low association in comparison to others. Also note that correlation with bi-tri-grams further enriches and distills information derived from monogram association. The labels derived from higher order phrases do not conflict with those obtained from single word association – in fact, they enlighten the use of supporting groups, possibly referring to unknown aspects of the objects (e.g. big, small as adjectives). We have shown, therefore, that using syntactic knowledge eliminates the problem we encountered in Section 3.1 regarding label learning. Also notice that, using syntactical knowledge (`box` for [**C**]) or association(`small` for [**SS**]) alone might lead to erroneous assignments.

While syntactic grouping lets us refine label learning, argument structure acquisition would assist in anaphora resolution and metaphor acquisition. We therefore discuss that in the following section. However, in the mean while, we would like to emphasize that the process of deriving syntactic information and mutual association of linguistic and perceptual elements 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.

4.3 Discovering Verb and Containment Structure

In the beginning of this section, we alluded to the advantages of discovering argument structure. Here, we describe plausible argument structures for previously grounded concepts of containment

and motion verbs. The rest of the paper would then deal with consequences of these structures in dealing with anaphors and metaphors. The following are the prominent IN-argument structures:

Pattern IN1

$$\left[\begin{array}{l} \\ \\ \\ the \rightarrow \\ \\ \\ \\ \end{array}\left[\begin{array}{c} big \\ bigger \\ empty \\ larger \\ little \\ small \\ smaller \end{array}\right]\right] \rightarrow \left[\begin{array}{c} \left[\begin{array}{c} circle \\ square \end{array}\right] \rightarrow \left[\begin{array}{c} goes \\ is \\ moved \end{array}\right] \\ \\ square \end{array}\right] \rightarrow inside \rightarrow the$$

Pattern IN2

$$\left[\left[\begin{array}{c} in \\ inside \\ into \end{array}\right] \rightarrow the\right] \rightarrow box$$

Pattern IN3

$$\left[\begin{array}{l} \\ circle \rightarrow \\ \\ \end{array}\left[\begin{array}{c} is \\ went \\ goes \end{array}\right]\right] \rightarrow \left[\begin{array}{c} in \\ inside \\ into \\ at \\ by \end{array}\right]$$

Pattern IN4

$$\left[\left[\left[\begin{array}{c} door \\ it \\ square \\ there \\ one \\ he \end{array}\right] \rightarrow \left[\begin{array}{c} is \\ went \\ goes \end{array}\right]\right] \rightarrow \left[\begin{array}{c} left \\ back \\ suck \end{array}\right] \rightarrow \left[\begin{array}{c} in \\ inside \end{array}\right]\right]$$

$$\left[\left[\left[\begin{array}{c} \\ \\ \\ \\ \\ \\ the \rightarrow \end{array}\right.\right.\right.\left.\left.\left.\begin{array}{c} ball \\ box \\ door \\ square \end{array}\right]\right]$$

From Fig 13(a), `in, into` and `inside` emerge as the dominant labels for the containment schema. They also belong to the same syntactic class (Patterns IN3 & IN4), discounting any syntactical conflict. Other monograms have very low associations, eliminating any chance of their being associated with the schema (not to mention their inclusion in separate equivalence groups). So, while monogram association itself unambiguously correlates perceptual and linguistic elements of containment, bi- and tri- gram associations, though syntactically ineligible (`in the box` is syntactically dissimilar to `in`, the label derived from monograms) to act as a label for containment, may provide further information. Perceptually, containment arises from interaction of trajector and landmark. The landmark, in spite of being an integral part of the background and hence the commentary, is undiscoverable through the attention model because it's never attended to explicitly (i.e. through gaze). The question we therefore want to address is, would syntactical information be capable in discovering label for the landmark (henceforth called the container or **[box]**, since it gives rise to containment)?

This, indeed, is possible through derivation of argument structure for IN. From the IN-patterns, the prominent structures that emerge are:

1. **[Group 1 word/phrase]** → **[IN]** → **[the]** (Pattern IN1)
2. **[Group 1 word/phrase]** → **[verb]** → **[IN]** → **[the]** (Pattern IN1)
3. **[Group 1 word/phrase]** → **[verb]** → **[IN]** (Pattern IN3)

4. **[Group 1 word/phrase]** → **[verb]** → **[other linguistic elements]** → **[IN]** (Pattern IN4)

5. **[IN]**→**[the]**→**[Group 1 word/phrase]** (Pattern IN2)

Argument structures can also be discovered for CHASE. The emergent structures are:

1. **[Group 1 word/phrase]** → **[CHASE (`chases/is chasing`)]** → **[Group 1 word/phrase]** (Fig **??**)

2. **[CHASE (`chased`)]** → **[by]** → **[the]** → **[little]** (Fig **??**)

3. **[CHASE (`chases`)]** → **[little]** → **[Group 1 word/phrase]** (Fig **??**)

derived from corresponding patterns:

Pattern C1

$$\left[\left[\left[\begin{array}{c} big \\ large \\ little \\ the \end{array}\right] \rightarrow \left[\begin{array}{c} box \\ square \end{array}\right]\right] \rightarrow \left[\begin{array}{c} chases \\ is \rightarrow chasing \end{array}\right]\right] \rightarrow \left[\left[the \rightarrow \left[\begin{array}{c} big \\ small \\ little \end{array}\right]\right] \rightarrow \left[\begin{array}{c} box \\ square \end{array}\right]\right]$$

Pattern C2

$$chased \rightarrow by \rightarrow the \rightarrow little$$

Pattern C3

$$\left[\begin{array}{c} chases \\ pushes \\ corners \\ the \end{array}\right] \rightarrow \left[little \rightarrow \left[\begin{array}{c} ball \\ block \\ box \\ circle \\ square \end{array}\right]\right]$$

The above are syntactical argument structures derived from linguistic input alone. We further correlate them to the *perceptual* argument structure to create grounded mappings. Consider the sentence " large square chases little square", which was uttered when **[BS]** was following **[SS]**, and the action features lied in Cluster 3 (chaser in focus, refer to Figure **??**). From the visual input,

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

Table 2: *Correlating perceptual and linguistic argument structure*

we therefore have the perceptual schema {[**BS**] CHASE [**SS**]}. From past experience, the agent is aware of the linguistic mappings for [**BS**], [**SS**] and CHASE, so that this correlation creates the mapping {`[chaser]` CHASE `[chased]`} → {[Group 1 phrase for [chaser]] [chase/chased/is chasing] [Group 1 phrase for [chaser]]}.[5] This is of course only one evidence. Table 2 shows statistics for all the sentences involving CHASE which fall into the discovered patterns. In the table, 'correct referent' means the perceptual and linguistic agents are not in conflict. 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. 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. Thus, the agent determines that with high probability, the construction for the action CHASE in English is `[chaser] chase+particle [chased]`.

The structures for IN show remarkable correlation in both linguistic and perceptual domain, thereby strengthening the acquisition of those structures. The structures also help acquire the label for the container or [**box**]. Consider the sentence "large square moves into the box". It adheres to the patterns 2 and 5 above. While this sentence is being uttered, the agent perceptually notices that [**BS**] is moving into the containment cluster. Therefore, the perceptual schema is: {[**BS**] IN [**box**]}. From the linguistic input, `large square` denotes [**BS**] and `into` denotes IN. So, [**box**] has two possible mappings in `moves` and `the box`. However, [**box**] is a physical object, and based on the past experience of the agent, should be assigned a linguistic element that syntactically confirms

5. Here `[chased]`, `[chaser]` are used by us for clarity - the system knows these based as a trajector-object distinction, in terms of visual focus

to concepts of other physical objects[6] like **[BS], [SS]** etc. , so that `box/the box` remains the only possible mapping. In fact, `inside/in/into the box` is not an isolated occurrence, but a prominent one through out the discourse(Fig 13(c)), thereby strengthening this mapping. This primary evidence is further strengthened from Table 2, where `box` has been taken as the 'correct' referent for **[box]** and with this assumption, 75% of the Structure 5 sentences agree in both perceptual and linguistic domain, thereby facilitating the acceptance of the assumption. In fact, of the 10 mismatched referents, only 1 is wrong (`in the door`, while the rest are due to unfamiliar referents (`in the room`, `in the square`).

4.4 A Brief Summary

In the work described above, we first started with a video and co-occurring commentaries to create a plausible grounded model for linguistic element acquisition. Objects, the actions they are involved in and containment through landmark were further schematized in unsupervised manner. They were then correlated with linguistic input to learn noun, verb and containment preposition labels and associate them with respective schemata. We then proceeded to derive syntactical structure from linguistic input only, and correlated them perceptually to create mappings between perceptual and linguistic argument structures. In summery, we have the following:

- a set of labels for objects, the actions they are involved in and containment
- a set of labels for the container involved in containment
- acquired sentential structure

We now proceed to handle anaphors and metaphors through these devices.

5. Anaphora: Recognition and Resolution

We have discussed how perceptual structures can be mapped to linguistic structures in the previous section. Through statistical evidence, we show how certain structures are accepted as favorable descriptions for the corresponding action schemas. We also allude to novel unfamiliar linguistic agents that appear in linguistic construction but are not mapped to any object in the perceptual domain. In this section, we argue that these anomalies are, in fact, the first evidences based on which

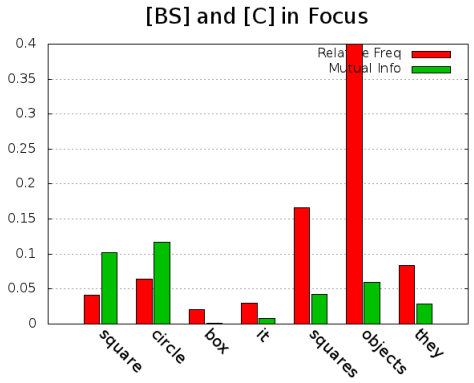
6. For the present set-up, all the moving objects and the container can be derived through image segmentation, thereby being similar 'physical' objects

	CHASE			IN			
Object	[**BS**]	[**SS**]	[**C**]	[**BS**]	[**SS**]	[**C**]	[**box**]
`it`	5	3	2	4	1	1	0
`he`	1	0	0	1	0	0	0
`room`	0	0	0	0	0	0	6
`block`	2	1	0	1	0	0	0
`ball`	0	0	0	0	0	5	0

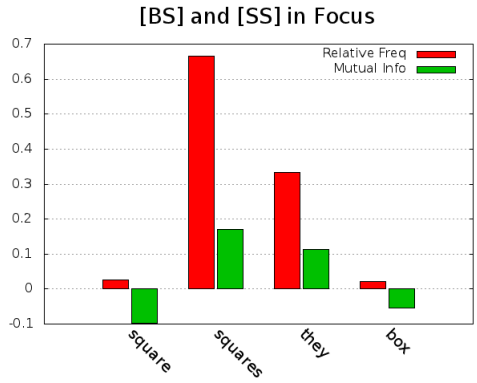
Table 3: *Perceptual to linguistic agent correlation for unassigned/unfamiliar Group 1 words.* Notice that `ball` and `room` are exclusively mapped to [**C**] and [**box**] respectively, thereby being treated as being synonymous with `circle` and `box`. `block` has been used for both [**BS**] and [**SS**], creating an equivalence with `square`. Linguistic elements like `it`, however, has been used to denote all three different trajectors, encouraging first signs of anaphora recognition.

the cognitive agent might acquire concepts of synonymy and anaphora. While attempting to discover synonyms and named entities of the discourse, the system discovers referentially stable mappings for fixed, single referents. But it also discovers several other units whose referents are dynamically determined by the recent discourse. This may be considered as a semantically-driven approach for discovering grammatical structures like 'the word order of arguments', and 'the phenomenon anaphora'.

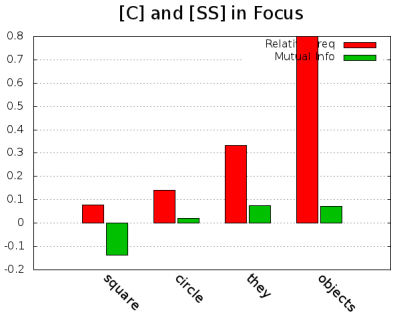
We motivate the process through an example. Suppose computing the relative motion features between the two objects in attentive focus, [**BS**] and [**SS**], the learner finds that the motion sequence matches the visual schema for the action CHASE, and given the order of the objects in the feature computation, the visual schema encodes the semantics of the predicate {[**BS**] CHASE [**SS**]}. Now consider sentences co-occurring with this action, e.g. `large square chases little square`. The agent recognizes the construction (CHASE Structure 1) and therefore, assigns `large square` to [chaser] and `small square` to [chased], as described in Section **??**. Now suppose it comes across the following: `large block chases little square`. Syntactically the sentence conforms to Structure 1, as `large block` is a Group 1 phrase. Assuming Grice's hypothesis that the speaker is providing necessary, sufficient and relevant information, with this evidence, the hearer would be prone to associate `large block` with [**BS**], since the two objects are the only ones involved in action and in visual salience, and the structure specifically administers this binding. it's possible that negative evidence might present itself once in a while: therefore, only those linguistic terms, which have statistically major evidence, will be acquired as synonyms while rare occurrences would be forgotten. the frequency estimates for novel Group 1 words, that conform to the learned structures for IN and



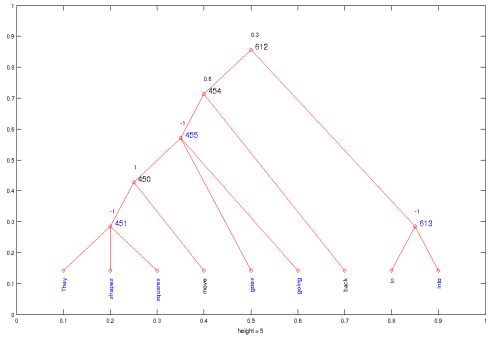
(a) Monogram association with **[BS]+[C]**



(b) Monogram association with **[BS]+[SS]**



(c) Monogram association with **[C]+[SS]**



(d) ADIOS structure for THEY

Figure 15: Figure showing association of single words with pairs of objects in the video. Both the association measures are shown. The red bars indicate the Relative Freq measure while the green bars are for the Mutual Information Measure.

CHASE, are shown in Table 3. The table lists linguistic elements and their perceptual mappings based on correlation between linguistic and perceptual argument structure. Notice that some words, like `room` and `ball` have been exclusively mapped to some perceptual agents, so that they are assimilated as labels for the corresponding objects ([**box**] and [**C**] respectively in this case). Some elements like `block`, have been assigned to objects of similar shapes ([**BS**] and [**SS**])[7], so that it can be considered similar to `square`.

7. The system can discover they are similar based on image processing for similar shapes, or from prior linguistic evidence that they have both been termed `square`.

Along with these, words like `it` appear, which are not assigned to any single entity but, can be applied to multiple referents, as is evident from Table 3. To the learner, this implies that this aspect, that these phrases can be applied to multiple referents, is stable, and not an artifact related to a single action or context. While other words like `the`, `and` etc. also appear in multiple contexts, the syntactic information derived from ADIOS assures that `it` is a Group 1 word, and therefore refers to an object in the context and isn't simply a linguistic construct. We take this as the first evidence of acquisition of the phenomenon of anaphora. The agent learns about some linguistic elements that are coreferent and also polysemous – they refer to different objects based on the context.

While the structural method we have followed is unable to discover any more anaphoric expressions like `it` due to lack of their use in CHASE and IN specific commentary (only 1 occurrence of `he` in both, two of `they`), the statistical association measure to discover noun labels can also discover them. In fact, we might notice that `it` appears prominently in Fig 9(a), though it's absent from others due to its very low frequency of occurrence in a large corpus. If, however, we consider utterances when two agents are simultaneously in attentional focus, `they` appears prominently. Fig 15 shows one word associations, using both of our previous measures, to pair of objects in focus; only Group 1 words have been considered, as the agent is now privy to that syntactical information. `squares` emerges as the dominant label for ([**BS**]+[**SS**]), whereas due to dissimilarity of shape, `objects` is the primary label for the other two pairs. `they`, however, appears as a prime contender in all the three, and the structure in Fig 15(d) confirms its syntactical eligibility. The point we would thus like to emphasize is that, linguistic elements that syntactically can be used for labels of objects and are used as such for multiple referents are easily discoverable for even a primitive agent. The regularity in their syntactical usage, the consistency of their mapping to the perceptual domain, and the assumption that utterances are informative assures the agent that those are regular features of language and not some domain-specific or syntax-only artifacts, thereby helping in their acquisition.

5.1 Further Discussion

In Section **??**, we raise several issues, like questions on how the first evidence of anaphoric elements are recognized, or zero-anaphoras are recognized. We have hoped to address the former in the preceding section. First evidence of anaphoras, we believe, comes from polysemous use of linguistic elements describing agents in action. The agent also notices that there are umpteen utterances in the discourse which provide only partial information of the perceptual actions, e.g. `chases the little`

`square`, while the perceptual action involves [**BS**] chasing [**SS**]. Only information about the [chased] is provided in the linguistic domain. Notice that the primary construction for CHASE assumes two agents – in fact, CHASE Structure 3 is essentially a part of Structure 1. This discrepancy between supposed mapping between perceptual and linguistic domains is characteristic of zero-anaphoras. The learner gradually accepts this seeming artifact as an aspect of language. There are two cues available to the early learner: a) that the relevant action involves two arguments, but fewer are available in the discourse, and b) that the missing argument refers to an antecedent in the discourse. In English, zero anaphora is a very common phenomenon. Even in our very small corpus, there are 570 agents, of which 99 are zero anaphors. Clearly this is a sufficiently high probability phenomenon which deserves the attention of the early learner. Once the absent argument is observed, it can be associated with the appropriate argument. Note that since this substitution is occurring at the semantic level and not in the syntax, only antecedents matching the activity will be considered.

The process of anaphora acquisition also motivates us for a way towards disambiguating agents in novel situations. The cognitive agent learns the existence of anaphoras through correlating perceptual to linguistic input, and the first resolution is from perceptual to linguistic domain, where perceptual agents are assigned linguistic concepts. However, later, when the agent comes across purely linguistic elements, it understands them in a grounded way(Langacker, 1987), where linguistic concepts are perceptually based. In fact, the recent proposal of Embodied Connectionist Grammar adheres to this idea: it proposes that as we come across linguistic utterances, we syntactically break them and try to create the context perceptually, in the mind. This, therefore, motivates for a 'world-view' approach to anaphora resolution. Sidner(Sidner, 1986) proposes that anaphor neither refers to another word nor co-refers to another word, but rather co-specifies a cognitive element *in the reader's mind*. Therefore, as a reader comes across linguistic input, it creates a context or world view in the mind, which is natural for it because it has acquired language through grounding it in perceptual space, a plausible method of which we have discussed in this work. So when it comes across anaphoric occurrence of `it` or `they`, according to the context it has in mind, it assigns the most appropriate referents. In fact, that's why most of the time anaphors refer to the most recent antecedent agent, since it's in attentional focus in the created perceptual domain. Estimating the probabilities in terms of frequencies even for this very small dataset, reveals that of the 99 zero anaphors, 96 refer to the most recent agent argument, often coming as a series e.g. `big square says ''uh uh, don't do that'' / pushes little square around /`

`pushes little square around again/ chases little square`. Thus, the most recent argument may emerge as a dominant reference pattern for zero anaphora. Also, we note how considerable knowledge beyond syntax is involved in the remaining situations e.g. `Door is shut/ Went into the corner`, which would be extremely hard to resolve through NLP techniques without a context.

6. Metaphor Acquisition: A Cognitive Approach

While there is ample evidence in literature to suggest that basic linguistic forms may be grounded(Roy & Reiter, 2005), only a small initial set of grounded linguistic concepts is needed. The majority of one's vocabulary is learned later purely from linguistic inputs(Bloom, 2000). Even metaphors are most likely learned from language usage only, without any physical grounding. The process by which this is achieved is not dealt with in this work, but it would seem that a similar physical experience can lead to different metaphorical mappings in different cultures. Given a large corpus that an early learner is being exposed to, it may learn that certain words share some co-occurrence statistics, forming natural clusters. As a quick example, searching for words that occur in similar contexts to 'love' in texts *Moby Dick* by Melville and *Sense and Sensibility* by Austen (using `text.similar()`[8] function in Python NLTK Toolkit), gives the following outputs:

- `Man Sea Ahab Air Bone Captain Chase Death Fear Hope Land`
- `Affection Heart Mother Town Dear Life Marianne Family`

Given such strong evidence for grouping of Objects/Entities (Man, Ahab, Captain) with emotions (*love, fear, hope*), an early learner, based on such word usage, along with an internal grounding for the Entity/Object class, may begin to impart object properties on feelings, leading to the glimmerings of something like the FEELINGS ARE OBJECTS metaphor.

Language is faced with the difficult task of having to handle the infinite variation of the world in terms of a much smaller number of constructions. A key process in language is to apply the same linguistic structures to different situations based on similarity. This is the primary argument in this work, which suggests that pre-linguistic concepts such as object categorization, containment, etc. - are used as the child tries to ground linguistic elements in terms of already acquired perceptual schemas. Therefore, nouns describing physical objects/substances seem to be easier to ground,

8. This function takes a word w, finds all contexts w_1ww_2, then finds all words w' that appear in the same context, i.e. $w_1w'w_2$

while abstract concepts are slower to acquire. Through exposure to common conversation, the child subsequently learns syntactic features/grammar of the language(Siskind, 1996). Consider an agent with this much information and competence. When the agent, by virtue of its knowledge of grammar, comes across sentences in which words describing emotions appear in the same context as a grounded object, it's but natural on a such a limited first evidence to impart grounded aspects of an object to this abstract concept. Later on, through more exposure, the agent might imbibe the abstract nature of the linguistic element.

To determine how far language usage alone can help shape the concept of metaphors, we compiled a list of sentences from (Lakoff & Johnson, 1980) and (Lakoff, Espenson, & Schwartz, 1991) that correspond to the ontological metaphor-mappings for Containers, Objects and Substances. *Ontological metaphors* are "ways of viewing events, activities, emotions, ideas, etc., as entities and substances"(Lakoff & Johnson, 1980). Our experiences with physical objects (especially our own bodies) provide the basis for an extraordinarily wide variety of ontological metaphors. The salient findings of this search were:

- Of the 85 sentences denoting Container metaphors, in 65, the abstract idea was imparted the image schema of a container based only on the prepositions *in/out*. In the rest 20, adjectives (*full, empty*) and verbs (*explode, erupt, fill*) took up the mantle.
- In all of the 63 sentences for Object metaphor, the Object property was imparted to the concept because VERB(A,B) took object arguments, i.e. verbs were the primary basis of metaphor mapping.
- Of the 42 sentences for Substance metaphors, 17 mappings were done based on adjectives (*more, less*) while the rest were of the type *Container contains Substance*, i.e. first the Container property was imparted, and then whatever was supposed to be inside the container was called a substance.

We observe that nouns, verbs and prepositions play a pivotal role in metaphorical mappings. But syntactic information alone may not be adequate in discovering rule-based metaphorical mappings; the possibility of over-generation is there at every step. Thus, for these mappings to reflect aspects of the human thought process, they would need to be grounded, an issue we have already handled in previous sections. While the grounding is restricted to a very small domain, we assume that similar grounding is possible in other domains to generalize the argument for a large corpus. With

this assumption, in the next phase, our system is exposed to broader linguistic corpus, and it encounters these symbolic units along with others for which also it is assumed to have similar primitive structures. By correlating among these structures based on the initial grounding that it has acquired earlier, it is able to enrich the primitive semantics of these units. By analysing the linguistic co-occurrences, it is able to learn *selectional preferences* from the corpus. In the rest of this section, we show how these selectional preferences lead to acquisition of clusters that suggest some of the metaphorical structures named by (Lakoff & Johnson, 1980).

6.1 Revisiting Argument Structure

In Section 4 we have detailed argument structure derivation for CHASE and IN. The discussion in the present section emphasizes the role of sentence structure in assigning similarity between different concepts, abstract or concrete. We want to investigate if the argument structures derived from the small discourse are extendable to a large general corpus. We used the Brown Corpus. We have shown how the placement of a concept as a container in IN schema imparts it the properties of a container, creating a Container metaphor map. We investigate if our primitive agent is capable of syntactically deriving such information about the container and the agents from a general corpus. The results of running ADIOS on Brown corpus are shown below:

IN-Trajectors 1.

$$the \rightarrow only \rightarrow \begin{bmatrix} college & survivors & time \\ thing & city & angel \\ junior & man & county \end{bmatrix} \rightarrow in \rightarrow the$$

2.

$$the \rightarrow \begin{bmatrix} mayhem & thermocouples & presumption & drama \\ two & wallpaper & dress & figure \\ girl & angel & men & variation \\ child & codes & intermediates & sailing \\ problem & woman & possibilities & fire \end{bmatrix} \rightarrow in \rightarrow the$$

3.

$$the \rightarrow other \begin{bmatrix} other \\ second \\ first \\ last \end{bmatrix} \rightarrow \begin{bmatrix} trial & place & blessing \\ participants & importance & things \\ action & year & men \end{bmatrix} \rightarrow in \rightarrow the$$

IN-Containers 1.

$$in \rightarrow the \rightarrow \begin{bmatrix} morning & thirties & beaker & night \\ house & ninth & spring & game \end{bmatrix} \rightarrow when \rightarrow the$$

2.

$$in \rightarrow the \rightarrow new \rightarrow \begin{bmatrix} ones & regions & army \\ government & land & stadium \end{bmatrix}$$

3.

$$in \rightarrow this \rightarrow \begin{bmatrix} field & recovery & car \\ flight & country & building \end{bmatrix} \rightarrow \text{“}$$

4.

$$in \rightarrow the \rightarrow same \rightarrow \begin{bmatrix} solution & manner & fashion & terms \\ international & community & way & slots \end{bmatrix}$$

While Patterns IN-Trajectors show that the algorithm has been able to isolate trajectors participating in the IN schema even from a diverse corpus, IN-Containers show detection of the containers. From the syntactical co-occurrence of {`college, time, county, museum`} with {`man, angel, survivors`} alone, the agent might begin assigning properties of a living object to inanimate landmarks and time. Similarly, the grouping of {`field, fight, recovery`} with {`car, building`} imparts physical container properties to those abstract concepts. These evidences, however, are isolated, and we therefore use selectional preference to statistically decide the prominence of such mappings in what follows. This gives the preliminary evidence that sentences can be broken

down into argument structures from purely statistical syntactic knowledge from the exposed corpus. CHASE occurs in less than 10 sentences in Brown, and that too, as a noun. So, due to lack of enough patterns, no argument structure can be found using the algorithm (although we get across this limitation in the following section).

6.2 Selectional Preference

We saw in Observation 2 that verbs play a major role in imparting metaphorical meanings. The said observation is also supported by (Shutova & Teufel., 2010), who claim that in 164 out of 241 metaphorical sentences, metaphoricity was introduced by verbs. Following discussion in the above paragraph, *given grammatical relations*, an agent should be able to find verbs similar in semantic or syntactic aspects to its repository of grounded forms. With the grounded concept of CHASE() and SPEND(), our system can find similar verbs based on their syntactic usage. Works that involve clustering concepts, using grammatical relations and lexical features, to capture their relatedness by association are abundant in literature(Shutova, Sun, & Korhonen, 2010). In fact, after Levin(Levin, 1993)'s semantic verb classification, supervised and unsupervised approaches(Korhonen & Marx., 2003) have been tried to automatically achieve the same goal. For this work, as we have only two grounded verb forms with us, instead of going for a clustering process, we use VerbNet(Schuler, 2005) to find other members of the semantic classes that CHASE() and SPEND() belong to. They are:

- CHASE, FOLLOW, PURSUE, SHADOW, TAIL, TRACK, TRAIL
- CONSUME, SPEND, WASTE

The reason we are going through this process is simple. As we saw in the two examples at the beginning of this section, verbs have a tendency to take up certain classes as objects. For example, WASTE() and SPEND() take up *money* and *time* as their objects most of the time. This brings *money* and *time* closer in the mind's cognitive space, leading to them being treated as similar concepts. While using a single verb to gauge the closeness of two concepts might incur error due to false evidence or unfair representation, using a class of semantically similar verbs would provide more credibility to the mappings, as there would be more evidence to support genuine mappings and less to support rare ones. Furthermore, the representational poverty of some verbs would be counterbalanced by inclusion of more frequent verbs of the same semantic class. In fact, with

inclusion of the CHASE-group, we now have a sizable number(774) of sentences to derive syntactic structures, which are presented below:

1.

$$the \rightarrow following \rightarrow \begin{bmatrix} comment & chairman & consideration & formula \\ reasons & thesis & breakdown & effects \\ indulgences & facts & information & relation \end{bmatrix} \rightarrow''$$

2.

$$[in \rightarrow the] \rightarrow following \begin{bmatrix} ways & paragraphs \\ manner & form \end{bmatrix}$$

3.

$$following \rightarrow the \rightarrow \begin{bmatrix} publication \\ balls \\ dispatch \\ example \\ practice \end{bmatrix} \rightarrow of$$

4.

$$the \rightarrow \begin{bmatrix} hush \\ decade \\ silence \\ year \end{bmatrix} \rightarrow that \rightarrow followed$$

The metric used the most in literature to measure regularities of a verb w.r.t. the semantic class of its argument (subjects, objects etc.) is **selectional preference (SP)**(Resnik, 1993). Some formulations of SP have been used previously for word-sense disambiguation(Resnik, 1993) and metaphor interpretation(Mason, 2004). While they have only been used for finding verb preferences, we will adapt them to include prepositional preferences too, so that we are able to learn more metaphors, especially containment metaphors, which will be otherwise hard to learn. In fact, including these verbs

lets us derive the argument structure for CHASE class, which was not possible before. The results are shown in Fig **??**. Notice that primarily three patterns emerge, {`following the` [AGENT]}, {`the following` [AGENT]} and {`the` [AGENT]`that followed`}. Also of interest is the pattern in Fig **??** – {`in the following` [AGENT]}, where [AGENT] is syntactically assigned both as the object of FOLLOW and IN, imparting it an Object and Container mapping (though such issues have not been considered in the grounded work in this write-up).

We follow the formulation presented in (Resnik, 1993). Suppose predicate p selects class c for the syntactic relation r, which we represent as *selects(p,r,c)*. For example, '*drink* takes *LIQUID* at object position' is represented as *selects(drink, object, LIQUID)*. The *selectional association*$(A(p,r,c))$ of class c for predicate p is then defined as:

$$S(p,r) = D(P(c|p,r)||P(c)) = \sum_c P(c|p,r) \log \frac{P(c|p,r)}{P(c)}$$

$$A(p,r,c) = \frac{1}{S(p,r)} P(c|p,r) \log \frac{P(c|p,r)}{P(c)}$$

While verbs have different syntactic relations like `verb-object`, `subject-verb` etc., the prepositions we are considering, have only one relation to the trailing noun, that of *Object of Preposition (***pobj***)*(Marie-Catherine de Marneffe & Manning, 2006). So, the formulation essentially remains the same and effects of the variable r are nullified.

We use WordNet(Feinerer & Hornik, 2011) as a knowledge-base for class c. WordNet was developed as a system that would be consistent with the knowledge acquired over the years about how human beings process language. Since an early learner, like our system here, would not have detailed information of all the synsets of a particular concept, we make use of only the lexical file types (25 in number), which encompass concepts like `quantity` and `possession`. These are the top level abstractions and we assume that an early learner is at a cognitive state where it has notions of these high level concepts.

6.3 Finding SPs

We used the Brown Corpus to test our model. All the sentences involving the grounded concepts were extracted. The sentences with prepositions were converted to the functional form of PREP(*pobj*) in a rather simple way: the first occurrence of a singular or mass noun(NN) after the preposition

in the tagged corpus was assigned to the concept. For example, the sentence fragment *into a hot cauldron* is converted to INTO(cauldron). [9] Handling sentences pertaining to the verb groups was tricky. The Stanford Parser(Marie-Catherine de Marneffe & Manning, 2006) was first used to extract VERB(*object*) relations. But owing to the large number of misclassification of dependencies, the dataset was rechecked by human annotators to correct discrepancies wherever present.

Following Resnik, $P(c|p,r) = freq(p,r,c)/freq(p,r)$ with

$$freq(p,r,c) = \sum_{w \epsilon c} \frac{count(p,r,w)}{classes(w)}$$

where $count(p,r,w)$ is the number of time w occurred, and $classes(w)$ is the number of classes it belongs to. To computationally manage a top-level WordNet semantic node, we approximated it over only those words which are part of the derived corpus.

6.4 Discovering Metaphor Mappings

The results of selectional association mapping are presented in Table 4.

For the SPEND() verb class, selectional association for classes `possession` and `time` are 0.790 and 1.1956 respectively. The representative nouns, in decreasing order of frequency are:

```
time(30), day(16), year(9), money(7), months(7)
```

This association strengthens the TIME IS POSSESSION metaphorical mapping. Since *money* is the largest contributor to `possession` in this context, it also leads to TIME IS MONEY metaphor. Also, this example shows why corpus only metaphor acquisition *without* a grounded world view is problematic. In (Mason, 2004), the class which has higher selectional association is considered a base class. However, in this case, the reverse happens. Here TIME is the target domain and MONEY is the source domain(Lakoff & Johnson, 1980). Therefore, source or target domain distinction can be done based on a world-view of grounded concept only.

The most occurring 'container' words were `world, way, order, years, case, states` etc. The resultant associations are shown in Table **??**. Notice that Location class has the highest association for container schema, activating a LOCATIONS ARE CONTAINERS mapping. Group class

9. One might notice that this technique for finding the head of the object of a preposition is problematic in the sense that a structure of the form *in iron cauldron* will be taken as *in iron* and not *in cauldron*. However, statistically, less than 1% of occurrences (in Brown corpus) are of the type where the immediate noun following the 'IN' is not associated with it. Furthermore, the Selectional Preference(SP) of 'IN' with 'IRON' over the whole corpus would be meagre, eliminating any effect of a miscalculated prepositional head.

CHASE		IN		SPEND	
Class	SA	Class	SA	Class	SA
communication	1.1959	location	0.6581	time	1.1956
act	0.3972	group	0.2007	possession	0.0790
time	0.3833	time	0.1754	quantity	0.0544
cognition	0.3692	cognition	0.1646	event	0.0362
event	0.1863	state	0.1455	Tops	0.0204
process	0.0625	artifact	0.0772	phenomenon	-0.0003
substance	0.0599	act	0.0579	motive	-0.0005
motive	0.0099	object	0.0545	shape	-0.0023
shape	0.0086	communication	0.0481	process	-0.0026
quantity	-0.0093	attribute	0.0322	feeling	-0.0035
animal	-0.0181	body	0.0244	animal	-0.0040
plant	-0.0198	shape	0.0196	relation	-0.0043
object	-0.0205	quantity	0.0119	food	-0.0051
phenomenon	-0.0259	phenomenon	0.0100	body	-0.0056
relation	-0.0270	feeling	0.0080	substance	-0.0063
feeling	-0.0301	relation	0.0066	object	-0.0077
food	-0.0365	Tops	0.0004	state	-0.0138
Tops	-0.0367	motive	-0.0007	attribute	-0.0159
body	-0.0576	process	-0.0042	location	-0.0176
possession	-0.0821	plant	-0.0154	group	-0.0324
location	-0.1036	possession	-0.0161	cognition	-0.0416
attribute	-0.1059	event	-0.0263	act	-0.0424
state	-0.1401	substance	-0.0269	communication	-0.0471
group	-0.1877	food	-0.0296	artifact	-0.0565
artifact	-0.2698	animal	-0.0420	person	-0.0755
person	-0.5019	person	-0.5342		

Table 4: Table showing selectional association(SA) of various class of words for containment (IN), CHASE and SPEND class verbs.

also has a strong association to containers, representative of the notion that groups or teams are visualized as containers (`in a group, in a team`). Time, Cognition and State also show high associativity, while Food and Person class demonstrate a significantly negative mapping. The negative numbers only represent the weakness of mapping, and should not be treated as repudiating existence of the same. The association measures only demonstrate that some mappings are used more in language, and consequently, are stronger in our cognition than others. In fact, in the original metaphor list(Lakoff & Johnson, 1980), the most prominent mappings to container are those of Cognition(15%), State(14%), Location(7.3%), Group(8.6%), Time(5.4%) and Act(4.8%), somewhat representative of their strength acquired from the whole corpus. Similarly, the least occurring classes in the list too are Plant(0.3%), Animal(0.3%) and Food(1%). Some examples of sentences from both the Brown corpus and metaphor list are presented below:

- STATE/COGNITION AS A CONTAINER
  - Meredith began falling in love. (Brown)
  - We're IN a mess. (Lakoff & Johnson, 1980)
- TIME AS A CONTAINER
  - We're well into the century. (Lakoff & Johnson, 1980)
  - There comes a time in the lives of most of us when we want to be alone. (Brown)
- LOCATION AS A CONTAINER
  - If you've travelled in Europe a time or two , . . . (Brown)
  - . . . and begin to feel at home in the capitals of Europe. . . (Brown)
- ACT AS A CONTAINER
  - How did you get into window-washing as a profession? (Lakoff & Johnson, 1980)
  - . . . the number of people travelling in your group. (Brown)

Similarly, CHASE shows the highest association with Communication, Act and Time. Given that CHASE takes objects in its argument structure, this association leads to the strengthening of the following metaphorical maps:

- COMMUNICATIONS ARE OBJECTS
  - . . . he headed rather than following the dictates of the Soviet Union .
  - If one follows the reports of the Congress , one finds that . . .
- ACTS ARE OBJECTS
  - Full payment of nursing home bills for up to 180 days following discharge from a hospital .
  - . . . Capitol Hill following the oath-taking ceremonies and ride down this historic ceremonial route .
- EVENTS ARE OBJECTS
  - a press conference following a meeting in the morning with Wagner
  - Show follows ceremonies
- TIME IS AN OBJECT
  - The event will be open to the public the following day .
  - The following week , I read in the Sunday paper that the students . . .

References

Bailey, D. (1997). A computational model of embodiment in the acquisition of action verbs..

Ballard, D. H., & Yu, C. (2003). A multimodal learning interface for word acquisition. In *Proc. of ICASSP*.

Bergen, B., Chang, N., & 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.

Bruner, J. (1983). *Child's Talk*. New York, NY:Norton.

Casasola, M., Cohen, L. B., & Chiarello, E. (2003). Six-month-old infant's cate- gorization of containment spatial relations. *Child Development*, *74*, 679–693.

Chomsky, N. (1957). *Syntactic Structures*. Mouton and co.

Chomsky, N. (1975). *Reflections on language*. Pantheon.

Dominey, P. F. (2005). Emergence of grammatical constructions: Evidence from simulation and grounded agent experiments author:..

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

Elman, J. L. (2005). Connectionist models of cognitive development: where next?. *Trends in Cognitive Sciences*, *9*(3), 111 – 117. ¡ce:title¿Special issue: Developmental cognitive neuroscience¡/ce:title¿.

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

Feinerer, I., & Hornik, K. (2011). *wordnet: WordNet Interface*. R package version 0.1-8.

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

G. Edwards, G. Ligozat, A. G. L. F. B. M., & Gold, C. M. (1996). A voronoi-based pivot representation of spatial concepts and its application to route descriptions expressed in natural language. In *In proc. 7th International Symposium on Spatial Data handling*, pp. 7B1–7B15.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279.

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

Iida, R., Yasuhara, M., & Tokunaga, T. (2011). Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 84–92, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441 – 480.

Kaplan, F., Oudeyer, P., & Bergen, B. (2008). Computational models in the debate over language learnability. *infant and child development*, *17*(1), 55–80.

Korhonen, A., Y. K., & Marx., Z. (2003). Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of ACL*.

Lakoff, G., Espenson, J., & Schwartz, A. (1991). *Master metaphor list: 2nd draft copy*.

Lakoff, G., & Johnson, M. (Eds.). (1980). *Metaphors We Live By*. University of Chicago Press.

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

Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press.

Locke, J. (1975). *An essay concerning human understanding*. Oxford, Clarendon.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271–296.

Maji, S., Singh, V. K., & Mukerjee, A. (2006). Confidence based updation of motion conspicuity in dynamic scenes. In *In Third Canadian Conference on Computer and Robot Vision CRV*.

Marie-Catherine de Marneffe, B. M., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Martin, B., & Tversky, B. (2003). Segmenting ambiguous events. In *Proceedings of the 25th annual meeting of the Cognitive Science Society*.

Mason, Z. J. (2004). Cormet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, *30*, 23–44.

Mukerjee, A., Vaghela, P. B., & Shreeniwas, V. (2004). Pre-linguistic verb acquisition from repeated language exposure for visual events. In *International Conference on Natural Language Processing*.

Mukerjee, A., Neema, K., & Nayak, S. (2011). Discovering coreference using image-grounded verb models. In *RANLP*, pp. 610–615.

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

Pereira, F. (2000). Formal grammar and information theory: Together again?. *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY*, *358*, 1239–1253.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA.

Prasov, Z., & Chai, J. Y. (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, pp. 20–29, New York, NY, USA. ACM.

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

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

Resnik, P. S. (1993). Selection and information: A class-based approach to lexical relationships. Tech. rep..

Rolls, E., et al. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, *9*(4), 467–480.

Roy, D., Hsiao, K., & Mavridis, N. (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(3), 1374–1383.

Roy, D. (2000). Integration of speech and vision using mutual information. In *Interna- tional Conference on Acoustics,Speech and Signal Processing(ICASSP00)*.

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

Sarkar, M. (2006). Grounded perceptual schemas: Developmental acquisition of spatial concepts. Master's thesis, IIT Kanpur.

Schuler, K. K. (2005). *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.

Shutova, E., Sun, L., & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proc. of COLING*, pp. 1002–1010.

Shutova, E., & Teufel., S. (2010). Metaphor corpus annotated for source - target domain mappings. In *Proceedings of LREC 2010*.

Sidner, C. (1986). Readings in natural language processing.. chap. Focusing in the comprehension of definite anaphora, pp. 363–394. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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

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

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11629–11634.

Spelke, E., & Hespos, S. (2002). Conceptual development in infancy: The case of containment. *Representation, memory, and development: Essays in honor of Jean Mandler*, 223–246.

Steels, L. (1997). Language learning and language contact. In *In Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pp. 11–24.

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

Steels, L., & Kaplan, F. (2001). Aibo's first words. the social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32.

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

Yuan, S., P. A. T. J., & Xu, F. (2011). Learning individual words and learning about words simultaneously. In *Proc. of the 30th Annual Meeting of the Cognitive Science Society*, pp. 3280–3285.