Discovering the concept of anaphora from grounded verb models

Kruti Neema and Amitabha Mukerjee
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur, Uttar Pradesh, India
krutineema@gmail.com, amit@cse.iitk.ac.in

***Abstract*—A number of computational models simulate the grounded learning of units of language in the early learner. But can this initial lexical knowledge be used to acquire complex grammatical notions such as anaphora? We build on earlier work, where we simulate a language learner with perceptual attention and learn, in an unsupervised manner, a set of action models along with the participating agents, and then the corresponding linguistic units (verbs). We consider how this knowledge may be used to bootstrap the learning of anaphora. Given an input video with moving shapes, the system considers human narratives that refer to this scene. The acquired perceptual schemas and their arguments are mapped to the appropriate verbs and nouns in the discourse. We first detect the synonyms of the arguments as the repeated labels used in the constructions referring to a known action scene. After ruling out the synonyms, we find that the anaphora remain as units that are referring to more than one grounded object. We show that both third-person singular and plural anaphors and even a common reciprocal anaphor ("each other") can be discovered. Finally, we show that in situations where the referent is missing altogether (*zero anaphora*), certain correlations may also be inferred from the regularities in the mapping between perceptual schemas and language.**

I. Introduction

The question of how a language learner may begin to comprehend, and eventually use, functional units of language, have been investigated computationally to a large extent [23], [7]. However, the question of how domain-general processes may help learning a grammar remains inadequately understood. Indeed, the question of learn-ability of grammar constitutes one of the most controversial areas in cognitive science. The "argument from the poverty of the stimulus" (term coined by Chomsky, 1980) [17] claims that any set of expressions of language can be explained by many grammars, so that the instances seen by the child are inadequate. This together with Gold's theorem [8] which shows that grammar is not learnable from only correct examples of language use, has bolstered the claims for nativism. Though the possibility that semantics may guide the learning of linguistic categories has been proposed [9], the mechanisms for this process are far from clear.

In this work we consider the computational feasibility of discovering the grammatical structure of anaphora by an early language learner, who has a few prior motivations and an inventory of machine learning algorithms, which it applies to discover regularities in its input, be it sensorimotor or linguistic. We consider this agent in a multi-modal interaction mode, where it is able to observe a scene that it can parse in terms of sensorimotor models of actions. Later it is exposed to linguistic utterances that refer to the scene. We show how such a system may acquire, in an unsupervised manner: a small set of a) perceptual action models, b) words in language for these actions and the participants, and c) constructions in language mapping these actions and participants. We then analyze a limited set of linguistic descriptors of actions in a 2D video, and demonstrate how such a system may be able to distinguish situations where a specific description is missing of a participant in a current action, and eventually correlate this missing description with the use of anaphora such as "it". We describe briefly these three inputs (they are described further in section III), and then the algorithm that is proposed. The subsequent sections detail the problem and the algorithm. Finally, we present some results for how such an algorithm works with this limited input.

Action models: While a number of grounded sensorimotor systems attempt to learn models for objects[19], [23], [14], models that consider actions, often use prior knowledge for visual parsing of actions [5], [7], [22]. Here we consider how an unsupervised process may acquire action structures from simple videos by clustering frequently observed sequences of motions. The model uses bottom-up (task-independent) dynamic attention [21] to identify the objects that are interacting. This work extends the results of Satish and Mukerjee [20], who consider two-agent spatial motions, and four cluster emerge; these turn out to correspond to the action categories [come closer], [move away], and two clusters corresponding to [chase]. These learned models or *image schemas* are acquired prior to language, and defined on the perceptual space. They are not related (as yet) to the linguistic input, though the latter may eventually come to modify it. The learned models include the agents participating in the action, which constitutes the cognitive arguments of the action.

Initial Vocabulary: Later, when our computational learner encounters language, it associates perceptual objects under attention to linguistic units in the co-occurring utterances. The strongest associations are learned as names for these objects (nouns) [14]. Next, it associates sentences uttered during the cognitive focus and correlates them with these actions. The strongest associations are learned as labels for actions (verbs) [20]. The actions [come closer] and [move away] do not have a very confident associations, but [chase] is strongly associated with the word "chase".

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

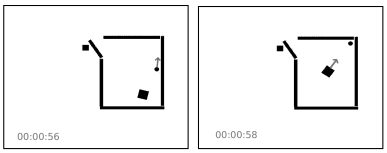

Fig. 1. *Multimodal input: 2D video "Chase"*: Three shapes, [big-square], [small-square] and [circle] interact playfully (velocities shown in gray). The system computes the relative velocities and positions and uses these as features to learn image schemas for some of the frequent actions. Later these action models are mapped to linguistic constructions to discover how arguments are being referred.

Linguistic Constructions: At this stage the system knows the names of the participants (e.g. "big square"), as well as the label for the action (e.g. "chase"). Among the utterances co-occurrent with the action, it now computes the probability of different orderings for the units (e.g. the ordering of "chase"+grammatical-particle, [chased] and [chaser]). Here [chased], [chaser] are used by us for clarity - the system knows this distinction only based on which one of the two objects it is primarily attending to. It determines that with high probability, the construction for the action [chase] in English is `[chaser] chase+particle [chased]`.

The question we consider in this work is how an agent who has come this far may now proceed to infer the presence of anaphoric mechanisms in grammar. The main insight is that while such a learner is aware of referentially stable mappings as with functional units, it can now discover that there are other units (such as pronominal anaphora) whose referents are dynamically determined by the recent discourse. Finally, we also consider how the learner may discover regularities where a referent is completely missing, i.e., the case of zero anaphora.

II. Recognizing Anaphora

Linguistic utterances often contain anaphoric references to antecedent objects in the discourse. There is considerable interest in computational models for disambiguating anaphora [13], but there seem to be less work on computational models for how a language learner can discover the phenomena of anaphora. While traditional syntactic theory proposes a number of solutions to the anaphora problem, a consensus seems to be emerging that modeling anaphora will require the combination of syntactic, semantic and pragmatic models [12]. Further, some anaphors involve deictic elements which transcend discourse. Rather than construct a grammar through syntactic analysis, this approach posits the discovery of these mechanism by correlating the linguistic expression with prior semantic knowledge in terms of sensorimotor schemas.

Here, we propose that an early learner with the three abilities given above may discover thatunlike regular linguistic units which refer to referentially stable categories, the anaphors are mapped to dynamic referents based on the recent discourse. The resulting models are grounded in the sensorimotor domain, and thus transcend syntax alone as a mechanism for modeling anaphora.

In this elementary demonstration of how this may work, we consider a simple 2D video. We analyze the motions of the agents, and correlate it with the linguistic narratives generated for these actions by humans. Anaphoric pronouns are frequently encountered in the narrations of the visual scene, 'it', and 'them' being the most popular in our experiments (e.g. *The big square is chasing it.*). Reciprocal anaphora also appear (e.g. *They are hitting each other*). Sometimes, an argument may be missing altogether, a phenomenon known as zero anaphora. This may occur due to the nature of spoken utterances which may otherwise be considered ungrammatical (*and chases the two*), or in conjoined phrases (*The big square moves out of the box and pushes the small square*).

In this work, we consider how a learner with the capabilities outlined above may use them to identify situations where the linguistic text uses the same unit (e.g. the third person pronominal 'it') to refer to multiple classes of objects. Such a model, by providing a bottom-up approach to the modeling of anaphora, may also resolve a number of issues in the current computational approach to anaphora in NLP literature. Such approaches are generally based on a variety of constraints and preferences, which are applied in different orders. These constraints are those discovered by the researchers, but how the early learner comes to acquire these remains unknown. The phenomenon of zero anaphora is even more complex, and has been considered in Chinese[4], Japanese[15], Spanish[6], but relatively little work is available in English. Here we propose a different, semantically grounded, approach towards modeling the phenomenon of anaphora.

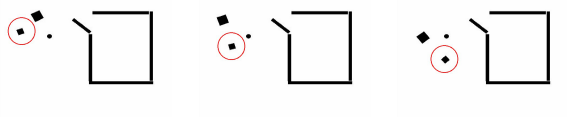

Fig. 2. Computed bottom-up attention windows during the part of the action [chase(big-square,small-square)]. The switching of attention between objects signals interaction between these objects and leads to the construction of 2-agent action templates or image schemas.

III. Grounded discovery of actions

The need for grounding words in language in terms of elements outside the symbolic system has been well-established [19]. We propose to ground our models of actions in the perceptual input, which is the major mode of infant learning in the first year [16]. As in [20], our input is an extremely simple 2D video with squares and circles moving around (a version of the well-known Heider and Simmel video [11]), and associated narratives[1]. We focus on two-agent action schemas, using a computational model of bottom-up attention [21] to identify actors in a specific interaction.

[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 [10].

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

The learner considers pairs of objects attended to within a short timespan, and computes two inner-product features for a) the relative-velocity and relative position and b) the relative pose and the sum of the velocities. The temporal histories of these feature vectors are then clustered using the temporal mining algorithm, Merge Neural Gas [24]. This work builds on the results of [20], where four action clusters are discovered, two of which correspond to [come-closer] and [move-away], and two others to [chase]. Chase has two clusters because it is asymmetric, and the primary attention may be on the chaser (cluster 1) or on the chased (cluster 2). By computing the feature vectors with the referents switched, the system can by itself determine this alternation.

Next, the computational learner attempts to map linguistic units to these clusters. For this, it first considers those sentences which overlap temporally with the period when the action clusters are active. One can align sentences with objects in attentive focus to identify the names of objects (nouns) [14], so at this point, we assume that the learner knows these nouns, 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 [come closer], [move away], and [chase] emerge as "move toward each", "move away", and "chase". Further, the learner is able to map certain linguistic constructions associated with the verb. Thus, it discovers that in the two clusters of chase, the sentences exhibit the construction `[chaser] chase+grammatical-marker [chased]` in 84% and 90% of the cases respectively. This construction matches sentences such as "The square chased the circle" or "The big square was chasing them". In a minority of cases, it also notes the construction `[chased] chase+particle by [chaser]`. We assume our computational learner has this level of competence before it attempts to detect substituted arguments and missing arguments in linguistic structures.

IV. Maps for Anaphora

In this work, we consider the problem of learning anaphora under two differing assumptions:

a. *Chase-only:* Consider only the action clusters discovered above. Since the linguistic forms for [move away] and [come closer] are very diffuse, we are restricted primarily to [chase]. However, we discover that [chase] also maps to the word "follow", and include sentences with "follow". Even then, our corpus of 35+15 sentences is very small, so the frequencies of specific strings are quite low.
b. *+Hit+Push* In the second model, we assume that in addition to [chase], we have action models and linguistic mappings for the actions [hit] and [push], which occur often in the commentary.

For the analysis, we consider the thirteen commentaries collected from U. Stanford students [10], as well as twenty-three collected from students at IIT Kanpur. These narratives have a wide range of linguistic variation - e.g. here are some sentences describing the sequence in figure 1:

a. Large square corners the little circle
b. Big square approaches little circle
c. Little square is moving away from the big square; and objects inside are moving closer together
d. Big block tries to go after little circle.

As can be seen, there are many synonyms for the nouns used in the narratives, and many perspectives highlighting different aspects of the scene. For our purpose, we are considering only those sentences where a target action (say [chase]) occurs concurrently with a narrative that contains the target verb ("chase"). Thus, the variations in perspective are not significant, though the noun synonyms are. However, we find that these synonyms can be detected, since the agent is aware of the actual objects and their roles in the action. After ruling out all such direct names, one observes that still many arguments of actions are missing in the linguistic expression. However, some terms, such as "it" or "they" may appear in the argument positions - these terms do not correspond to a single object, but appear to be applied to different objects in different situations. This is the beginning of the discovery of an anaphora model. The language learner categorizes the actions and with the help of the grounded verb-models, determines the objects that should act as the arguments of these actions. On matching this with the given sentence, it is found that anaphoric references are resolved. Furthermore, on many occasions, arguments are completely missing. These will lead our computational learner towards the phenomenon of zero anaphors.

A. Algorithm

Algorithm 1 A plausible approach towards the discovery of anaphora.

Input :

- Set of timestamped action predicates *Verb*(*arg1*, *arg2*)
- Set of timestamped narrative sentences

Alignment :

1. Align co-occurrent predicates with sentences containing the corresponding verb.
2. Increment the object associations against each language phrases L_i::
   - For linguistic constructs of the form, $\langle L_1 \rangle$ verb $\langle L_2 \rangle$ map L_1 to arg1 and L_2 to arg2
   - For constructs of the form, $\langle L_1 \rangle$ verb by $\langle L_2 \rangle$ map L_1 to arg2 and L_1 to arg1
3. For set of three agents, plus pairs (total 6 object-groups), estimate the conditional probability P(object/language phrase).
4. If the probability is close to 1, the language phrase is likely to be a proper synonym of the corresponding object.
5. If some linguistic units are acting as a synonym for multiple objects, their referent may not be fixed, but may depend on some other aspect.

Step 5 in algorithm 1 gives the first indication of phenomena such as anaphora. Furthermore, in many cases, some action

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

arguments may be altogether missing. Such situations may eventually suggest the presence of zero anaphora.

B. Working: Pronominal anaphora ("it")

Consider the action shown in fig. 3, which is matched with the image schemas acquired by the learner. In these frames, only two objects are in attentive focus (fig. 2, the big square ([BS]) and the small square ([SS]). Computing the relative motion features between these two, the learner finds that the motion sequence matches the image schema for the action [chase], and given the order of the objects in the feature computation, it obtains the predicate *chase([BS], [SS])*. Now, consider the sentence *large square chases little square*, whose timestamps overlap this action predicate. Matching the arguments with the linguistic construction, it is able to associate "large square" with [BS] and "little square" with [SS]. Now, "big square" and "little square are already known as labels for [BS] and [SS] [14], so "large square" is associated with [BS] as a possible synonym map.

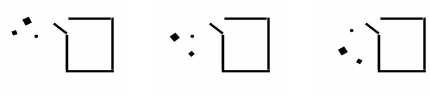

Fig. 3. Frame sequence in video concurrent with the utterances *large square chases little square*, *it is chasing the small box* and *chases little square*. The motions in the video matches the image schema [chase], with [BS] and [SS] as the participants. These co-occurring language phrases are then mapped to the predicate *chase(BS,SS)*

Another sentence aligned with the same action, *it is chasing the small box* results in the associations "it":[BS], and "small box":[SS]. Note that extremely frequent words like "the" are dropped in this analysis. Similarly, in *chases little block*, there is no referent at all for [BS], and "little block" is identified as a possible synonym for [SS].

C. Estimating probabilities for Action maps

The computational results in the following make a number of assumptions which are circular - in the sense that we assume some things are known which may strictly become known only later. For example, in [20], the mutual exclusivity principle was used - i.e. if we have a name for concept 1, the same word may not apply for concept 2. This was used to rule out known nominals when looking for verbs. But though our work is based on this, in the following, we shall be finding lots of synonyms, which clearly violate mutual exclusivity. Similarly, we will assume that frequent words like "the" are being dropped, though strictly, this should also apply to words like "it" which we are considering. Thus, the computations shown next should be taken as indicative of the plausibility of the approach, and not as a very formal implementation of the algorithm.

Further, since the data is very limited, we shall be presenting results under two differing assumptions: a) using only the image schema for [chase], and b) assuming that similar image schemas can be discovered for other motion verbs such as [hit] and [push] (these have not been learned by us).

Furthermore, the chase action that is discovered by the unsupervised learning only captures [chase] when the objects are near each other as in fig. 2. An important chase sequence, occurring towards the end of the video (fig. 4) happens where the chaser is on the other side of the box. This sequence is not detected by the image schema learnt by chase. Hence these sentences are not aligned with the chase action and are dropped in the computation under assumption 1.

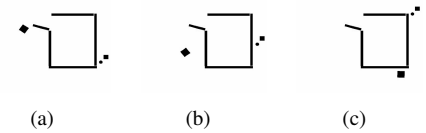

Fig. 4. Frame sequence concurrent with the chase action involving the three objects, towards the end of the video.

1) Assumption 1: Discourses mapping [chase] only: Of the three classes of actions for which we have acquired image schemas from the perceptual data, the narratives for [come-closer] and [move-away] have widely varying constructions. Focusing on the action chase, we discover that it maps to two verbs in the linguistic descriptions: "chase", and "follow". Constructions for both have the structure `[chaser] verb+particle [chased]`.

Using the mechanism illustrated in the example above, we can formulate associations to discover synonyms (algorithm 1 step 4). There are only 36 + 9 sentences with "chase" + "follow", so the data for these arguments is rather sparse. After ruling out phrases that have a sample size of one, cases where the conditional probability of the entity given the phrase is 1, is taken as a synonyms (names known earlier in italics):

BS : *big square*, square, big box, large square, big block, bigger square
SS : *little square*, small square, little box
C : *circle*, little circle, ball, small circle

Now, after ruling out synonyms and infrequent phrases (those occurring only once), we are left with three units - "it", "them" and "each other" (table 5). We were surprised ourselves that all three instances found are anaphora. After many such experiences, the system notices that these units do not have a fixed referent, and hence it searches for other regularities by which their referents can be identified. This may be the start of a process which leads to the idea of anaphora.

2) Assumption 2: + [hit] + [push]: While we have no computational models for actions such as [hit] and [push], there is considerable evidence that these concepts are typically acquired fairly early, and also reflected in early vocabularies [3]. In the analysis next (table 6), we assume the availability of [hit] and [push] models in addition to [chase], and consider the same analysis as above, but now on the larger set of sentences encoding these actions. A few additional synonyms are learned ("he" for [BS], "small box", "little block" for [SS]). Also the labels "square and circle", and "little circle and square" are

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

Phrase	# Phrase	BS /Phrase	SS /Phrase	C /Phrase	BS&SS /Phrase	SS&C /Phrase
it	10	0.5	0.4	0.1	0	0
them	5	0	0	0	0.2	0.8
each other	3	0	0	0	0.66	0.33
[missing]	15	0.46	0.2	0.33	0	0

Fig. 5. Conditional probability computation (with values in the column headers) for the non-synonymical arguments in sentences mapping [chase] action.

associated with the combination [SS&C], sentences mapping multiple predicates where both were involved in a patient role. These results may also be interpreted as a slightly advanced stage for the learner, when it has acquired these additional structures.

The remaining words are not assigned to any single entity but as in the [chase]-only case, they can be applied to multiple referents. 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. The learner may now attempt to discover other regularities in how the referents for each of these words is assigned. This requires even greater vocabulary, since the prior referent must also be known.

Phrase	# Phrase	BS /Phrase	SS /Phrase	C /Phrase	BS&SS /Phrase	SS&C /Phrase
it	19	0.63	0.26	0.11		0
each other	10	0	0	0	0.9	0.1
they	6	0	0	0	0.66	0.33
them	5	0	0	0	0.2	0.8
[missing]	29	0.59	0.24	0.17		0

Fig. 6. Conditional probability computation (with values in column headers) for the arguments of [chase], [hit] and [push].

Focusing on the word "it", and assuming a greater inventory of verbs, we can consider sequences of sentences such as *The bigger square just went inside the box / Looks like it is chasing the small square*. The "it" in the second sentence is known to our learner as [BS] based on the video parse, and one notes how the agent in the previous sentence is also [BS]. In another situation we have *The large square was chasing the other square / And it got away*. Here the "it" refers to the most recent antecedent, [SS] (though in other examples, it refers to the parallel antecedent). In the chase-only case, we note that "it" refers to the immediately previous referent in 6/10 situations. Two cases involve plural vs single disambiguation: e.g. *Big square is chasing them / They outrun it*, and one case involves parallel reference, e.g. *Now the big square is hitting the small square / It has hit it again* (in fact, unlike our learner, the reader may have difficulty disambiguate the "it"s here). While the referent identification pattern isn't very clear, the learner realizes that "it" at least refers to some earlier referent in the discourse.

Further, even reciprocal anaphors such as "each other" can be recognized since sentences such as *they hit each other* overlap with multiple predicates with switched arguments (*hit([BS],[SS])* and *hit([SS],[BS])*). Beyond this little domain, as our learner is exposed to thousands of linguistic fragments every day, these regularities are likely to get reinforced.

Finally, considering the cases of missing arguments, 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. 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*.

V. Conclusion

We have outlined how an unsupervised approach correlating prior sensorimotor knowledge with linguistic structures, might be used to eventually learn complex aspects of grammar such as anaphora. Here, the learner first acquires an inventory of sensorimotor image schemas that model commonly seen actions. These schemas include not only symbolic aspects such as the list of agents involved in the action, location, manner, but also the subsymbolic aspects such as the nature of the motion. Focusing on the action arguments, which are apparent in the perceptual model, we show how their absence in the linguistic expression can be detected. Arguments that are not mentioned by a direct name are shown to have high correlation with units such as the third person pronominal "it", the accusative "them", and also the reciprocal anaphora "each other". Also, we highlight many cases of zero anaphora, and show how these may also be inferred, most commonly as the most recent agent in the scene.

This work is of course, merely a start. While some of the action models, e.g. [chase], have been instantiated in earlier work, that model covers only a limited range of the set of actions covered by "chase" in English. Also, it is by no means clear that other action models needed for such a step can be similarly learned. However, there is considerable work that hints at the infants being able to use perceptual cues to learn

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

the base model of many motion primitives of this nature [16]. Also, the demonstration system is based on a simple 2D video.

A second lacuna is the necessity that the perceptual input be available to the agent in the immediate context. This is perhaps needed for the very earliest learners who are our focus here. However, for more mature speakers, anaphora is handled mentally, in terms of the context created by previous discourses. While we do not explicitly consider this situation, it may be argued that the image schemas invoked here by direct perception, are actually capable of being simulated, and thus the same schemas may be activated during comprehension of symbolic structures. This also provides a mechanism for modeling the cognitive structures during discourse comprehension, and perhaps the same phenomenon of missing arguments would be detected there also. Thus, once the mechanisms of anaphora is grasped in a grounded manner, most likely through some degree of multimodal input, the purely linguistic phenomenon, and other abstractions leading from it, may also be enabled.

Although this demonstration is rather limited, it does highlight several points. First, it underscores the role of concept argument structures in aligning with linguistic expressions. It provides some evidence for the position that some aspects of semantics may be ontologically prior to syntax, at least for human-like learning processes. Of course, once language is acquired it modifies these early semantic structures in profound ways that we do not begin to consider here.

Secondly, it addresses the very vexed question of learning grammar from domain-general capabilities. While a computational demonstration such as this cannot provide full answers, certainly it raises a very plausible mechanism, and attempts to learn some complex grammatical constructs such as anaphora.

Finally, it addresses some of the issues related to learning language from shared perception, such as the radical translation argument highlighted by Quine's *gavagai* example [18], and instantiates a possibility, first highlighted by [2], that dynamic attention may prune the visual input and align with linguistic focus.

For all humans, the vast majority of our vocabularies are learned later purely from the linguistic input [1]. But this is only possible because of the grounded nature of the first few concepts, without which these later concepts cannot be grounded. Thus the perceptually grounded nature of the very first concepts are crucial to subsequent compositions. This paper takes the argument one step further by suggesting that this perceptual grounding may also be key to learning of other grammatical phenomena such as anaphora.

The demonstration here clearly opens many more questions than it answers, but the purpose of this paper is to provide an initial argument for a semantically-grounded approach to discovering grammar. The intent is to argue that this line of investigation may be worth pursuing, and to provide a straw model where a number of questions can be asked. It will take considerable work before we know whether such an approach would scale for other types of anaphora related phenomenon, and if so, for what kinds of input, and if at all these reflect models that are consistent with infant learning modalities.

REFERENCES

[1] P. BLOOM, *How Children Learn the Meanings of Words*, MIT Press, Cambridge, MA, 2000.

[2] J. BRUNER, *The ontogenesis of speech acts*, Journal of child language, 2 (1975), pp. 1–19.

[3] E. V. CLARK, *First language acquisition*, Cambridge University Press, 2003.

[4] Y. CUI, Q. HU, H. PAN, AND J. HU, *Zero anaphora resolution in chinese discourse*, in CICLing, 2006, pp. 245–248.

[5] P. F. DOMINEY AND J.-D. BOUCHER, *Learning to talk about events from narrated video in a construction grammar framework*, Artificial Intelligence, 167 (2005), pp. 31–61.

[6] A. FERRÁNDEZ AND J. PERAL, *A computational approach to zero-pronouns in spanish*, in In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000, 2000.

[7] M. FLEISCHMAN AND D. ROY, *Why verbs are harder to learn than nouns: Initial insights from a computational model of intention recognition in situated word learning*, in Proceedings of the 27th Annual Meeting of the Cognitive Science Society, 2005.

[8] E. M. GOLD, *Language identification in the limit*, Information and Control, 10 (1967), pp. 447–474.

[9] A. E. GOLDBERG, D. CASENHISER, AND N. SETHURAMAN, *Learning argument structure generalizations*, Cognitive Linguistics, 14 (2004), pp. 289–316.

[10] B. HARD, B. TVERSKY, AND D. LANG, *Making sense of abstract events: Building event schemas*, Memory & cognition, 34 (2006).

[11] F. HEIDER AND M. SIMMEL, *An experimental study of apparent behavior*, 57 (1944), pp. 243–59.

[12] Y. HUANG, *Anaphora: a cross-linguistic approach*, Oxford University Press, 2000.

[13] R. MITKOV, *Anaphora Resolution*, Oxford University Press, 2003, pp. 267–283.

[14] A. MUKERJEE AND M. SARKAR, *Perceptual theory of mind: An intermediary between visual salience and noun verb acquisition*, 2006.

[15] T. NOMOTO AND Y. NITTA, *Resolving zero anaphora in japanese*, in Proceedings of the 6th conference on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, 1993.

[16] R. M. G. PASEK, KATHY HIRSH, ed., *Action meets word: how children learn verbs*, Oxford University Press US, 2006.

[17] G. K. PULLUM AND B. C. SCHOLZ, *Empirical assessment of stimulus poverty arguments*, The Linguistic Review, 19 (2002), pp. 9–50.

[18] W. V. O. QUINE, *Word and Object*, MIT Press, Cambridge,MA, 1960.

[19] D. ROY AND E. REITER, *Connecting language to the world*, Artificial Intelligence: Special Issue on Connecting Language to the World, 167 (2005), p. 112.

[20] G. SATISH AND A. MUKERJEE, *Acquiring linguistic argument structure from multimodal input using attentive focus*, aug. 2008, pp. 43 –48.

[21] V. K. SINGH, S. MAJI, AND A. MUKERJEE, *Confidence based updation of motion conspicuity in dynamic scenes*, in CRV '06: Proceedings of The 3rd Canadian Conference on Computer and Robot Vision, Washington, DC, USA, 2006, IEEE Computer Society, p. 13.

[22] J. M. SISKIND, *Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic*, Journal of Artificial Intelligence Research, 15 (2001), pp. 31–90.

[23] L. STEELS, *Evolving grounded communication for robots*, Trends in Cognitive Sciences, 7 (2003), pp. 308–312.

[24] M. STRICKERT AND B. HAMMER, *Merge SOM for temporal data*, Neurocomputing, 64 (2005), pp. 39–71.

978-1-4244-6902-4/10/$26.00 ©2010 IEEE