

Why Verbs are Harder to Learn than Nouns: Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning

Michael Fleischman (mbf@mit.edu)

Deb Roy (dkroy@media.mit.edu)

Cognitive Machines Group

The Media Laboratory

Massachusetts Institute of Technology

Abstract

We present a computational model that uses intention recognition as a basis for situated word learning. In an initial experiment, the model acquired a lexicon from situated natural language collected from human participants interacting in a virtual game environment. Similar to child language learning, the model learns nouns faster than verbs. In the model, this is due to inherent ambiguities in mapping verbs to inferred intentional structures. Since children must overcome similar ambiguities, the model provides a possible explanation for learning patterns in children.

Introduction

A growing trend in the cognitive sciences is to model how words are grounded in sensory-motor representations, providing explanations for how words map to the physical world (see Roy, in press; Reiter & Roy, in press, and references within). The social aspects of word meanings, however, are not addressed in these physically-grounded approaches. Although researchers have emphasized the central role of social factors in language acquisition (Tomasello, 2001), virtually all computational models to date have ignored this perspective (Regier, 2003).

We present a preliminary computational model of language acquisition that addresses aspects of social inference. The model highlights the role of intention recognition in word learning by formalizing the conceptual structure of intentional action. Further, it suggests that children's slower learning of verbs than nouns may partially be due to structural ambiguities inherent in intention inference.

Much research on language acquisition has sought explanations for the asymmetry between noun and verb acquisition in the developing cognitive or linguistic abilities of language learners (Gentner, 1982; Snedeker and Gleitman, 2004). Such work rightfully assumes that something in the nature of verbs makes them inherently more difficult to learn than nouns. Gillette et al. (1999) qualify this assumption by showing that the ability of subjects to learn a word in a human simulation paradigm is highly correlated with the "concreteness" of that word, and further, that verbs are considered less concrete, or

perceivable, than nouns. This intuition of concreteness, however, is not well defined and, as Snedeker and Gleitman (2004) discuss, what makes one class of referent more perceivable than another is unclear. Gleitman (1990) gives a number of examples showing why verbs may be considered less perceivable than nouns, such as observational equivalences between verbs (e.g. "chase" and "flee") and differences in temporal persistence between objects and actions.

In this work we posit two kinds of structural ambiguity in the perception of actions that explain the difficulty in learning perceptually grounded verbs. We present a formalization of the conceptual structure of intentional action and use it to model aspects of social understanding in verb learning. Rather than learn mappings directly from words to observations, the model posits an intermediate step that infers hidden intentional structures based on observed events. The model learns mappings from words to observed events and objects in these inferred intentional structures. We train the model using data collected from pairs of human participants interacting in a shared virtual game environment in which one person uses unconstrained speech to guide the actions of the other. Results indicate that the model does indeed learn nouns faster than verbs, and further, suggest that this is because, even while an intentional action can be interpreted in multiple ways, the objects involved in that action often remain stable.

The Ambiguity of Intentional Action

While the ambiguity associated with describing actions has been studied extensively (Vallecher and Wagner, 1987; Woodward et al., 2001; Gleitman, 1990), few researchers have proposed computational models of actions that explain those ambiguities. To motivate the model, consider an example of trying to learn words for actions taken by a player in a videogame world such as that shown in Figure 1. Assume the player must respond to spoken requests by performing various tasks in order to win the game. Further, assume that a language learner observes these interactions (verbal requests paired with the player's actions situated in the virtual world). As a

language learner, one hears the novel word “grok” uttered and observes the player mouse-click on the leftmost door. Now, based only on sensory observation, a number of possible interpretations of the word are possible. “Grok” may mean open the door, or alternatively, move to the door. Or perhaps “grok” is a command to let another player into the room, or for the player to go find some needed object (such as an axe).

Such situations demonstrate two distinct types of ambiguity for intentional action, which we represent as a lattice in Figure 2. The leaf nodes represent physical observations of actions, while the root nodes represent the highest-level intentions behind those actions.

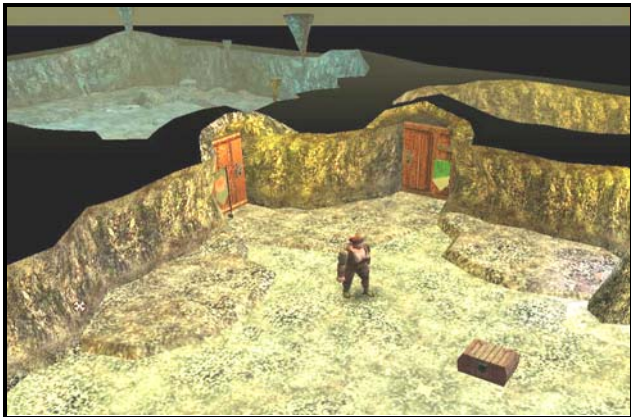


Figure 1: Screen shot of shared virtual environment.

The first type of ambiguity shown here, which we refer to as a *vertical ambiguity*, describes the ambiguity between the *find axe* versus *open door* interpretations of “grok.” Here the ambiguity is based on the level of description that the speaker intends to convey. Thus, given that the speaker did intend to look for the axe, if questioned about their action they would answer “yes” to *both* of the questions: “Did you mean go find the axe?” and “Did you mean open the door?”

The second type of ambiguity, referred to as *horizontal ambiguity* describes the difference in interpretation between the *find axe* versus *let in player* meanings of “grok.” In this case, the high level action behind the sensed action is ambiguous. Unlike with vertical ambiguities, only one of these actions is typically intended. Thus, if the speaker were questioned about their action, they could answer in the affirmative to *only one* of the questions: “Did you mean let another player in?” and “Did you mean go find the axe?”¹

By representing intentional actions as a lattice, both vertical and horizontal ambiguities are captured. This representation serves as the foundation for incorporating intention recognition, an important aspect of social understanding, into our model of word learning.

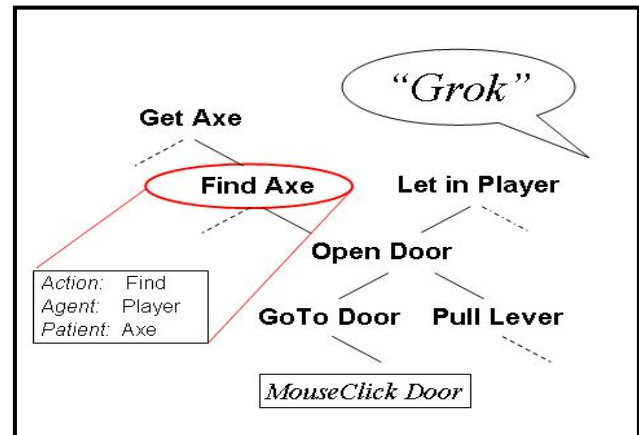


Figure 2: Graphical representation of intentional action. Two distinct ambiguities surrounding actions are represented by the horizontal and vertical dimensions of the lattice of semantic frames.

Intention Recognition

Intention recognition is the ability to infer the reasons underlying an agent’s behavior based on a sequence of their observed actions. We develop a probabilistic context free grammar (PCFG) of behaviors that allows for the building of intention lattices in much the same way that a PCFG for syntax allows for the parsing of sentences (e.g., Collins, 1999). This idea of a “grammar of behavior” dates back at least to Miller et al. (1960) and has been suggested more recently by Baldwin & Baird (2001). In our formulation, the grammar consists of *intention rules* that describe how an agent’s high level intentional actions (e.g., *find axe*) can lead to sequences of lower level intentional actions (e.g. *open door, go through door, open chest*). Such rules mirror syntactic rules where high level syntactic categories produce lower level categories (e.g. NP → DT N).

Unlike syntactic rules, however, each node of an intention lattice encodes a semantic frame that contains the participants of the action and their thematic roles (actor, patient, object, etc.). For example, in Figure 2 (see insert), the node labeled *find axe*, comprises a frame with a FIND action, a PLAYER agent, and an AXE patient. In this initial work, the intention rules are created by hand (see below). We acknowledge that learning such rules automatically must be a focus of future work in order to scale our modeling approach.

By formalizing the grammar of behavior as a PCFG, we can treat intention recognition as a parsing problem over observed actions (as in work on plan recognition, e.g. Pynadath, 1999). We borrow established algorithms used for syntactic parsing in computational linguistics (Stolke, 1994). However, instead of parsing words in a sentence, we parse observed actions of an agent in an environment.

¹ While vertical ambiguities may have a parallel in objects (e.g. animal-dog-poodle) (Rosch, 1976) horizontal ambiguities are unique to intentional actions.

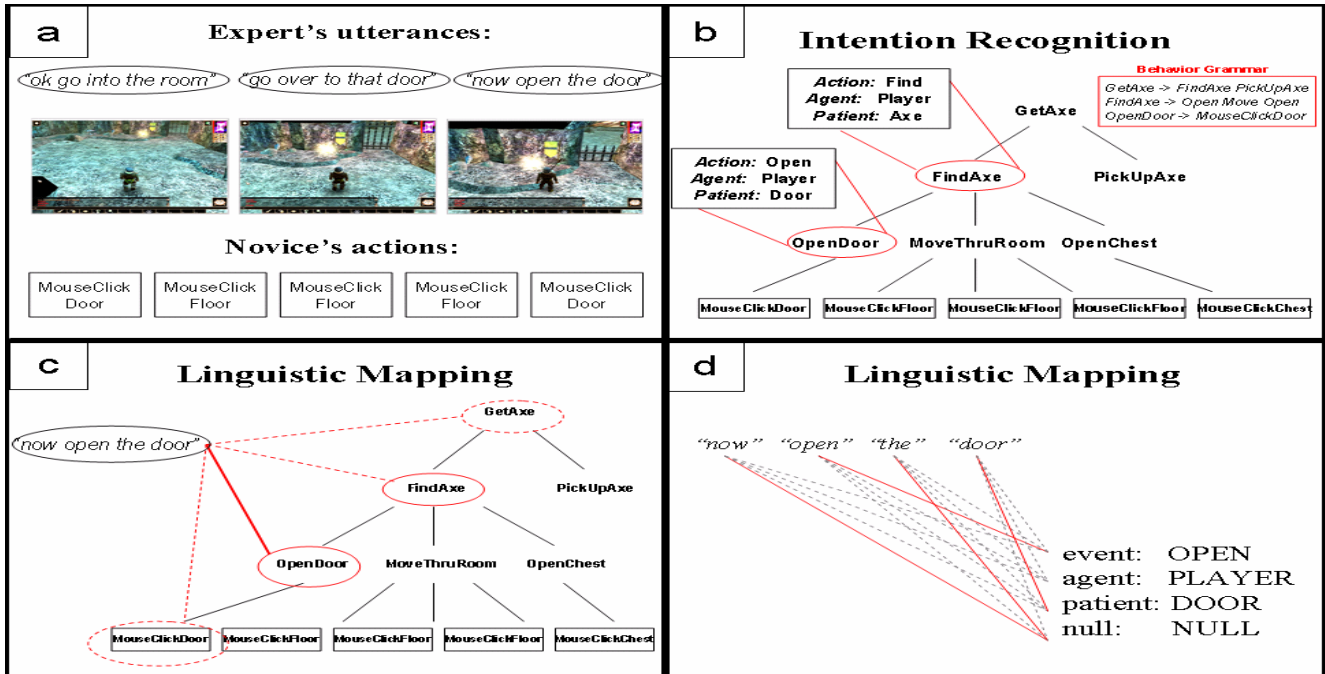


Figure 3. a) Parallel sequences of speech and actions are recorded from subjects as the expert guides the novice through a virtual environment. b) An intentional tree is inferred over the novice's sequence of observed actions using a probabilistic context free grammar of behaviors. This most likely parse given the PCFG resolves the horizontal ambiguity inherent in intentional action. Each node in the tree is a different level of intentional action and is encoded by a semantic frame. c) The vertical path from a leaf node in the tree (i.e. observed action) to the root (i.e. highest order intentional action) contains multiple possible levels of intention to which an utterance may refer. Linguistic mapping uses d) the Expectation Maximization algorithm to estimate the conditional probabilities of words given roles to resolve this vertical ambiguity.

Thus, as each action is observed, the parsing algorithm infers a lattice that describes all of the possible higher order intentional actions that could have produced it. As more and more actions are observed, the algorithm focuses in on a single most likely sequence of higher order intentional actions. For any sequence of actions that composes a completed game, the algorithm finds the single most likely intentional *tree* that has the most likely higher order intentional action as the root node and has each observed action as a leaf node (see Figure 3b). By finding the most likely intentional tree, the algorithm resolves the horizontal ambiguity surrounding a sequence of observed actions.

For a language learner observing such a sequence, the inferred tree can be seen as the conceptual scaffolding onto which utterances describing those events are mapped. In these initial experiments, the temporal alignment between a spoken utterance and the observed action to which it corresponds is hand annotated (a focus of future work is the relaxation of this assumption). Even given this annotation, and the most likely intentional tree for a sequence, there still remains the question of what level of description the speaker had in mind for their utterance, i.e. the vertical ambiguity.

We represent this ambiguity using the *vertical path* from the root of the most likely tree to the leaf node temporally aligned to the target utterance. The vertical ambiguity is thus represented by the multiple nodes that

the given utterance could refer to along this vertical path (see Figure 3c). To resolve this ambiguity we turn to the linguistic mapping procedure described below.

Linguistic Mapping

As described in the previous section, each node in an inferred intention lattice consists of a semantic frame. Our linguistic mapping algorithm attempts to learn the associations between words in utterances and the role fillers in these frames. We represent these mappings as the conditional probabilities of words given role fillers [i.e. $p(\text{word} \mid \text{role filler})$]. By formalizing mappings in this way, we can equate the problem of learning word meanings to one of finding the maximum likelihood estimate of a conditional probability distribution.

The Expectation Maximization (EM) algorithm has been used to estimate such distributions in many applications (e.g. Brown et al., 1993). EM takes a set of parallel inputs and finds a locally optimal conditional probability distribution by iterating between an Estimation (E) step and a Maximization (M) step. In our model, input consists of a sequence of utterances, each paired with a set of semantic frames (the nodes from the corresponding vertical path).

To understand our use of EM, let us first assume that we know which node in the vertical path is associated with an utterance (i.e., no vertical ambiguity). In the E

step, an initial conditional probability distribution is used to collect expected counts of how often a word in an utterance appears with a role filler in its paired semantic frame (see Figure 3d). In the M step, these expected counts are used to calculate a new conditional probability distribution. By making the one-to-many assumption that each word in an utterance is generated by only one role filler in the parallel frame (but that each role filler can generate multiple words) the algorithm is guaranteed to converge to the maximum likelihood estimation of the conditional distribution after multiple iterations of the E and M steps. Following Brown et al. (1993), we add a NULL role filler to each semantic frame which acts as a “garbage collector,” generating common words that don’t easily map to objects or actions (e.g., “the,” “now,” “ok,” etc.).

The above procedure describes an ideal situation in which one knows which semantic frame from the associated vertical path should be paired with a given utterance. As described above, this is not the case for language learners who, even knowing the intention behind an action, are faced with a vertical ambiguity as to what level of description an utterance was meant to refer (shown in Figure 3c).

We extend the EM algorithm to account for this vertical ambiguity by creating an outer loop that iterates over all possible pairings of utterances and semantic frames along the vertical path. For each of these possible pairings, standard EM is run and a conditional probability distribution is estimated. After all pairings have been examined, their estimated distributions are merged together, each one weighted by their likelihood. This procedure (detailed in Figure 4) continues until a stopping criterion based on cross-validation performance is reached. The utterance/frame pair with the highest likelihood is thus the most probable resolution of the vertical ambiguity.

- | |
|--|
| <ol style="list-style-type: none"> 1) set uniform likelihoods for all utterance/frame pairings 2) for each pair, run standard EM 3) merge output distributions of EM (weighting each by the likelihood of the pairing) 4) use merged distribution to recalculate likelihoods of all utterance/frame pairings 5) goto step 2 |
|--|

Figure 4. Extended EM used in to resolve vertical ambiguities.

Representing linguistic mappings as conditional probabilities not only allows us to apply efficient algorithms to the task of word learning, but also leads to a Bayesian formulation of language understanding in which understanding an utterance is equivalent to finding the most likely meaning (i.e. semantic frame) given that utterance (Epstein, 1996):

$$p(\text{meaning} | \text{utterance}) \approx p(\text{utterance} | \text{meaning}) \cdot p(\text{meaning}) \quad (1)$$

These posterior and prior probabilities have natural analogues to our representations of linguistic mapping and intention recognition. Specifically, the posterior $p(\text{utterance} | \text{meaning})$ can be estimated by the probability of the most likely alignment of words to role fillers (using the probabilities described in this section). Further, the prior $p(\text{meaning})$ can be estimated by the probability of the most likely inferred intentional tree (i.e. the probability given by the by the PCFG parser, as described previously).

Model Evaluation

Data Collection

We developed a virtual environment based on the multi-user videogame *Neverwinter Nights* (Figure 2).² This software includes an authoring tool enabling creation of new games within the virtual environment. A game was designed in which a human player must navigate their way through a cavernous world, collecting specific objects, in order to escape. Subjects were paired such that one, the *novice*, would control the virtual character, while the other, the *expert*, guided her through the world via spoken instructions. While the expert could say anything in order to tell the novice where to go and what to do, the novice was instructed not to speak, but only to follow the commands of the expert. The purpose behind these restrictions was to elicit free and spontaneous speech that is only constrained by the nature of the task.

The subjects in the data collection were university graduate and undergraduate students (8 male, 4 female). Subjects were staggered such that the novice in one trial became the expert in the next. The game was instrumented so that all the experts’ speech and all of the novices’ actions were recorded during game play. Figure 3a shows example screen shots of a game in progress along with the two associated parallel sequences of data: the expert’s speech and novice’s actions.

The expert’s speech is automatically segmented into utterances based on pause structure and then manually transcribed. The novice’s action sequences are parsed using a hand built behavior grammar to infer a tree representation of the novice’s intentions (see Figure 3b). In the current experiments, the entire sequence of actions composing a game trial is parsed at once and linguistic mapping is performed using the most likely tree from that parse. This batch processing allows for much more reliable intentional trees (since all of the actions in a game have been observed), but must be relaxed in future work to more accurately simulate a human learner’s limited temporal window into the world.

In hand building the behavior grammar, two sets of rules were created: one to describe agents’ possible paths of movement and one to describe non-motion actions. The movement rules were built semi-automatically, by

² <http://nwn.bioware.com/>

enumerating all possible paths between target rooms in the game. The action rules were designed based on the rules of the game in order to match the actions that players must take to win (e.g. opening doors, taking objects, interacting with non-player characters, etc.). Rules were built and refined in an iterative manner, in order to insure that all subject trials could be parsed. Because of limited data, generalization of the rules to held-out data was not examined. Probabilities were set using the frequency of occurrence of the rules on the training data. A major focus of future work will be the automation of this process, which would bring together the inter-related problems of language acquisition and task learning.

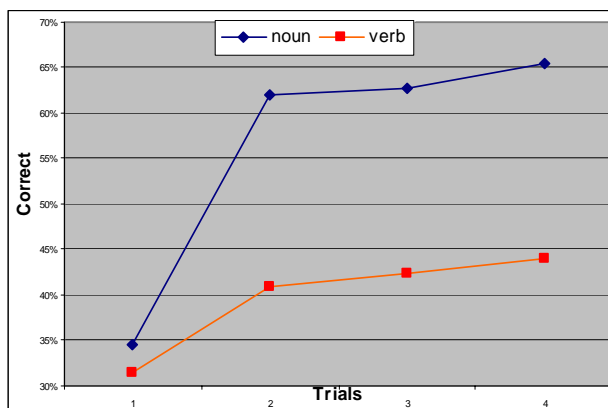


Figure 5: Comparison of noun and verb learning. Accuracy of the model for nouns vs. verbs is as a function of the number of trials used in training.

Having collected the utterances and parsed the actions, the two streams are fed into the extended EM algorithm, where the semantic roles from the novice's intention tree are mapped to the words in the expert's utterances. By iterating through all possible mappings, the algorithm eventually converges to a probability distribution that maximizes the likelihood of the data (see Figure 3c-d).

Experiments

To evaluate the model, the linguistic mapping algorithms are trained on the first four trials of game play for each subject pair and tested on the final trial. This gives on average 130 utterances of training data and 30 utterances of testing data per pair. For each pair, the number of iterations, beam search, and other initialization parameters (see Moore, 2003) are optimized using data from all *other* subjects.

For each utterance in the test data, the likelihood that it was generated by each possible frame is calculated. We select the maximum likelihood frame as the system's hypothesized meaning for the test utterance, and examine how often the system maps each word of that utterance to the correct semantic role. Word mapping accuracies are separated by word class (i.e. nouns and verbs) and

compared. Further, the amount of training data is varied to simulate the effect of experience on language learning.

Results

Figure 5 shows the performance of the model as a function of the number of trials used for training. The figure shows the model's understanding accuracy for both nouns and verbs, and indicates that, although initially the verb classes are learned with similarly poor accuracy, by the second trial there is a large jump in performance on nouns over verbs. Table 1 shows the model's accuracy on the 10 most frequent nouns and verbs from the test data, along with the frequency of those words in the training and test data. The Table shows that the model's accuracy for nouns is significantly ($p < 0.01$) greater than its accuracy for verbs, even though fewer nouns than verbs were present in training. Further, the Table shows a greater amount of variation in the performance on verbs.

Discussion

Results show that the model follows the trend in human learners to favor noun learning in the early stages of development. The model of course makes vast simplifications compared to actual human language acquisition. We believe, however, that the causes underlying the model's different learning rates for nouns and verbs provide useful insights into human learning. These results cannot be explained by frequency differences in the training data, in which more verbs appeared than nouns. Rather, the results follow directly from the model's formalization of the conceptual structure of intentional action and the inherent ambiguity of those actions.

To see this, we can decompose the notion of perception into sensation plus interpretation. The model does not give any preference to objects over actions in terms of sensation. To the model, objects and actions are both just role fillers in semantic frames. There is a difference, however, when considering interpretation.

As described above, the interpretation of action sequences introduces two distinct kinds of ambiguity. Such interpretations are represented in the model as different nodes in an intention tree, where each node represents a different semantic frame to which an utterance may map. The key to our noun/verb asymmetric result lies in the fact that, while each node of an intentional tree (i.e. semantic frame) has a different action role, often the object roles in different levels are the same.

For example, in Figure 2, the actions FIND, OPEN, and MOVE occur only once along the vertical path from root to leaf. However, the object DOOR occurs multiple times along that same path. In a word learning scenario, this means that even if the model misinterprets what level of intention an utterance describes, because object roles are repeated at multiple levels, it still has a good chance of

mapping the nouns in the utterance to their correct roles. However, because action roles are more specific to their level of description, if the model misinterprets the level, linguistic mapping for the verb cannot succeed.

This pattern is consistent in the data used for training, where, for each vertical path in an intention tree, the same action role is seen on average 1.05 times, while the same object role is seen 2.05 times. Thus, it is the ambiguity of actions and the recurrence of objects in a vertical path, which causes the model to learn verbs slower than nouns.³

Table 1: Word level results of model. The testing accuracy, and frequency in training, for the ten most frequent nouns and verbs in the test set are presented.

Word	VERBS			Word	NOUNS		
	Frequency	Accuracy	Test (%)		Frequency	Accuracy	Test (%)
	Train	Test			Train	Test	
Go	342	69	73.9	door	249	47	85.1
Get	96	16	6.3	chest	89	22	54.5
Open	65	17	23.5	portal	49	10	80.0
take	59	14	50.0	key	33	9	100
bash	47	12	66.7	axe	29	11	54.5
follow	31	6	33.3	password	28	7	100
talk	28	7	0.0	lockpick	26	6	100
Turn	22	6	0.0	diamond	25	7	71.4
Ask	19	7	42.9	lever	24	7	85.7
Teleport	10	4	0.0	archway	23	5	100
ALL	719	158	44.0	ALL	575	130	65.2

Conclusion

The primary contribution of our model is the use of situational context to support intention recognition in word learning. This work is one of the first steps we are aware of that introduces social considerations into a computational model of word learning (also see Yu et al., 2003). The experimental results we present, while very preliminary in nature, mirror the noun/verb asymmetry seen in human language development. Further, the model provides an explanation of the phenomenon by means of its formalization of the conceptual structure of intentional action.

In future work, the model will be extended to address the role of syntax in word learning, focusing particularly on how a formalization of syntactic bootstrapping (Snedeker & Gleitman, 2004) relates to the vertical and horizontal ambiguities of intentional action. Further, we will examine how behavior grammars can be automatically learned as language is acquired. Finally, experiments will be conducted in which the model is compared to human subjects performing comparable language learning tasks.

³ While formalizing object ambiguity may dilute this effect. Research on “basic level” descriptions (Rosch, 1976) suggests that ambiguity for objects may be different than for actions.

Acknowledgments

Peter Gorniak developed the software to capture data from the videogame used in our experiments.

References

- Baldwin, D. & J. Baird (2001). Discerning Intentions in Dynamic Human Action. *TICS*. 5(4).
- Brown, P. F. Della Pietra, V. J. Della Pietra S. A. & Mercer., R. L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics* 19(2).
- Collins, M. (1999), Head-Driven Statistical Models for Natural Language Parsing. PhD Dissertation, University of Pennsylvania.
- Epstein, M. (1996) Statistical Source Channel Models for Natural Language Understanding Ph. D. thesis, New York University.
- Gentner. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj, editor, *Language development: Vol. 2. Language, cognition, and culture*. Erlbaum, Hillsdale, NJ, 1982.
- Gillette, J., Gleitman, H., Gleitman, L., Lederer, A. (1999). Human simulation of vocabulary learning. *Cognition*, **73**.
- Gleitman, L. (1990) The structural sources of verb meanings. *Language Acquisition*, 1(1)
- Miller, G. A., Galanter, E. and Pribram K. H. (1960). *Plans and the Structure of Behavior*. New York: Halt.
- Moore, Robert C. 2004. Improving IBM Word Alignment Model 1. in *Proc. of 42nd ACL*, Barcelona, Spain.
- Pynadath, D. (1999). Probabilistic Grammars for Plan Recognition. Ph.D. Thesis, University of Michigan.
- Regier, T.. (2003) Emergent constraints on word-learning: A computational review. *TICS*, 7, 263-268.
- Rosch, E., C.B. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem.(1976) *Basic objects in natural categories*. *Cogn. Psychol.*, 8.
- Reiter, E. and Roy. D. (in press). Connecting Language to the World. Special issue of Artificial Intelligence.
- Roy, D.. (in press). "Grounding Words in Perception and Action: Insights from Computational Models". *TICS*.
- Snedeker, J. & Gleitman, L. (2004). Why it is hard to label our concepts. To appear in Hall & Waxman (eds.), *Weaving a Lexicon*. Cambridge, MA: MIT Press
- Stolcke., A. (1994) Bayesian Learning of Probabilistic Language Models. Ph.d., UC Berkeley.
- Tomasello, M. (2001). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language Acquisition and Conceptual Development*. Cambridge University Press.
- Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological Review*, 94, 3-15.
- Woodward, A., J. Sommerville and J. Guajardo (2001), How infants make sense of intentional action. In Malle, Moses, Baldwin *Intention and Intentionality*. MIT Press.
- Chen Yu, Dana H. Ballard and Richard N. Aslin (2003), “The Role of Embodied Intention in Early Lexical Acquisition”, Proceedings of the Twenty-Fifth Annual Meeting of Cognitive Science Society. Boston, MA.