Perceived Affordances as a Substrate for Linguistic Concepts

Peter Gorniak (pgorniak@media.mit.edu)
Deb Roy (dkroy@media.mit.edu)
MIT Media Laboratory
Cambridge, MA 02139 USA

Abstract

We propose the use of perceived affordances to computationally model linguistic concepts in situated language use. Perceived affordances are mental structures that capture the relationship between the language user and the embedding situation. We employ computational models of human perceived affordances to understand situated language: language that dynamically depends on the current physical environment and the goals and plans of communication partners. To support this theory of situated language understanding, we describe an implemented system that understands verbal commands situated in a virtual gaming environment. The implementation uses probabilistic hierarchical plan recognition to generate perceived affordances. We have evaluated the system on its ability to correctly interpret free-form spontaneous verbal commands recorded from unrehearsed game play between human players, and find that it is able to "step into the shoes" of human players and correctly respond to a broad range of verbal commands in which linguistic meaning depends on social and physical context.

Introduction

We have recently introduced the theory of Affordance-Based Concepts (ABCs) (Gorniak, 2005; Gorniak and Roy, 2006). This theory has at its core the intentional link between language users and the world treating predicted interactions as the basic building block for conceptual representation. Other theories often limit themselves to specifying the structure of concepts as opposed to how concepts come to be about the world. To tightly couple the internal structure of concepts with their intentional and functional use, the theory proposes that each element of a concept must make a prediction about the world, thus crossing over from the mind to the world. Every concept thus becomes both a property of the language using system, and of its relation to the embedding world. These structural elements are called *perceived affordances*, yielding a theory of *Affordance-Based Concepts*.

In this paper, we introduce a computational model that employs plan recognition as a mechanism for finding and ranking the perceived affordances of a person engaged in co-operative tasks. Situated language interpretation is modeled as a process of filtering perceived affordances. In effect, the complete meaning of linguistic expressions is only understood when words are meshed with the situation in which they are used. To evaluate the model, we describe an implementation of the model that interprets situated language collected from people playing a multiplayer computer game. This implementation is based on a probabilistic, hierarchical plan recognizer in the form of an Earley parser, To understand the game players' language, a syntactic parser filters the predictions (perceived affordances) this plan recognizer makes at the time of an utterance. Language understanding thus is cast as a filtering process on perceived affordances, which are elements that naturally span the boundary between the language user's mind (by virtue of being subjective predictions taking into account the language user's goals) and the world (by virtue of taking into account the structure of the world). Our results show that this implementation can accurately predict how human players respond to spoken commands issued by their human partners, demonstrating the viability of our approach for modeling the interpretation of context-dependent language on the basis of perceived affordances.

Related Work

There exist a number of competing or complementary theories of concepts. Laurence and Margolis as well as Prinz give overviews of the proposals and the debates surrounding them (Laurence and Margolis, 1999; Prinz, 2002). Many theories of concepts adopt a fundamental bias: that the use of concepts by an actual language using system, and thus their connection to the world, is secondary to their internal structure and formal properties. This bias is exhibited most strongly by the traditional definitional theories. Where they acknowledge a link to the world at all, for example in the form of observation sentences (Carnap, 1932), the link is portrayed as simple and passive. In fact, the only function usually considered is that of categorization, a feature shared with Prototype Theory (Rosch, 1975). Closest in spirit to our proposal here is perhaps Theory-Theory (Carey, 1985), but only in so far as it acknowledges the importance of viewing concepts as embedded in a structure used to reason about particular domains. While Theory-Theory emphasizes internal mental theories, however, it leaves unspecified their connection to the world except, again, in the case of categorization where it explains a human tendency towards essentialism. All these theories thus largely deal with internal mental structures and at most tackle the problem of categorization using these mental structures. Categorization is only one possible use of concepts, however, and arguably not the most basic one. Concepts are intimately tied to the goals of the conceptualizer and thus a theory of concepts should not only explain categorization behaviour, but more importantly cover intentional aspects of concepts such as the predictions they imply, and generally their use in achieving goals.

In a similar vein, the computational theories and systems addressing the problem of how a word comes to be about the world, treat this problem as one of adding sensory categorization functionality to a definitional theory of concepts (Harnad, 1990). There are a number of systems that have been built according to this paradigm, including Regier (1996)'s work on learning words for spatial relations, Plunkett et al. (1992)'s image labelling investigations, Yu, Ballard, and Aslins model of word learning Yu et al. (2003), as well as some of our own previous models of word grounding (Roy et al., 2002; Roy, 2002). These systems often address additional aspects of this type of grounding, such as how categories are learned by the language user (Regier, 1996; Roy, 2003) and how concepts grounded in this way can be combined (Gorniak and Roy, 2004). We have previously presented work on understanding situated language using plan recognition (Gorniak and Roy, 2005a,b), but that work separated the linguistic understanding from the plan recognition, whereas they are tightly coupled here. Few sensory grounded language systems can be found that can act autonomously using the concepts they maintain. Where action is involved, at hand is usually either once more a sensory categorization problem (Siskind, 2001). Some systems, such as KARMA, do understand language in terms of action representations, and these are perhaps most related to the work presented here. So far, however, these systems have not been shown to understand situated language dependent on a dynamically evolving situation (Narayanan, 1997). There exists work on computationally modelling affordances more abstractly as a theoretical tool to explore linguistic mechanisms (Steedman, 2002), as well as in a non-linguistic setting to model a robot's interactions with the real world (Stoytchev, 2005). Our robotics work has led Roy to propose a theory for grounding linguistic concepts in physical interaction (Roy, 2005). That work complements that presented here as a proposal for linguistic meaning based on interactions with the world at a far more detailed and fine grained level of physical (sensory-motor) experience than considered here. Finally, Fleischman and Roy (2005) have explored the word learning problem using an intention recognition framework very similar to the one presented here, but focusing on the difference between learning nouns and verbs.

Theory

The term *affordance* was coined by Gibson (1977). Rather than focusing on image-like representations that are similar to, or correspond to, the light information impinging on the retina, he proposed that perception encodes what the external world affords the perceiver. Thus, extended surfaces are perceived to provide support for walking on, if the surface is of an appropriate size relative to the perceiver and sturdy enough to hold the perceiver's weight, and the perceiver is actually able to walk. Perceived affordances, those affordances mentally represented by a perceiver, fulfill our requirements of a representation: they are the product of perception of the world, they encode some aspect of the structure of the world relative to the perceiver, and they predict a possible interaction between perceiver and world. By implying a prediction, they can be falsified.

In our context, a perceived affordance encodes some aspect of the state of the world, as well as a prediction made based upon this state. The world's state might be represented by encoding a fragment of the world's history, as is common in computational models. The perceived affordance's prediction may be representationally explicit, such as a list of possible ways to pick up a cup, or it may be implicit, such as an encoding of the cup's geometry together with a model of possible hand movements and configurations. A perceived affordance addresses the possible action prediction problem at a single level of representation. The possible ways to pick up a cup versus the choice of possible breakfast foods one faces every morning are on very different levels of representation. They are connected, however, in that a possible breakfast choice may include pouring a cup of milk, and thus picking up a cup. To make mental representation feasible it is important to keep these levels of affordances related yet distinct. Keeping them distinct allows one to reason on a single level, to achieve more concise yet still approximately Markovian state encodings and to employ the representation and reasoning methods that are best for that level. Keeping them loosely connected, on the other hand, allows for predictions that span levels and lets one fill in the details of high level plans, creating a hierarchy of perceived affordances.

Note that so far we have not invoked the notion of objects per se – perceived affordances are about the structure of the world that can be exploited to make predictions. This structure can be below the level of everyday objects, for example when it concerns the geometry of a graspable surface, which may or may not be part of a larger structure that we usually label "doorknob." Having replaced the notion of objects with the notion of structural elements called affordances, we can now re-introduce objects as *bundles of affordances*. A doorknob yields a set of interactions, as determined by its physical properties and the agent's abilities. When we engage in an active process of representation to distinguish objects within the structure of the world, we carve out a set of local affordances in the world and consider it an object. This process is not arbitrary, however, as it exploits the pre-existing structure of the world, including our own abilities. Thus while concepts of objects are the product of our perception, representation and actions, and while we may decide to cut up the world into different sets of objects at different times, we are externally constrained in our object categorizations by our own structure and that of our environment.

In the following studies this unified representation of objects as bundles of perceived affordances lets us capture the situation in terms of its possible functions for the agent. For example, a door is represented by the uses an agent might have for it, such as unlocking it, opening it or walking through it. This lets us interpret language in a representational substrate that already includes predicted actions and abstractions and thus turns understanding into a filtering process on this substrate. For example, "open the door" selects a subset of the perceived affordances of the listener in his or her present situation that involve opening available doors.

Concepts of objects are instances of the more general class of structures we call concepts. Each concept is a bundle of perceived affordances. In addition to representing concrete everyday objects, concepts can represent sets of structures in the world not limited to a single agent and object. Allowing arbitrary bundles of affordances gives the Affordance Based Concept theory a unique representational power, but the use of affordances imposes limits as it is subject to constraints from the structure of subject and environment. One aspect of this power is the ability to represent abstraction. For example, the command "let me into the next room" in our studies selects a more abstract interaction of changing rooms that at lower levels expands out into the listener unlocking or destroying the door to the next room, or pulling the correct lever to open it, followed by the speaker moving to the next room. This is also an example of concept composition in which the filter functions of lexical items are combined during the linguistic parsing process. Thus, while "room" selects sets of affordances available in any single room in the virtual environment, "next room" selects only those requiring exactly one room change by the speaker.

Implementation

In the implemented system, the structure of perceived affordances hinges on the notion of a hierarchical plan. A plan is a sequence of one or more steps an agent takes or considers taking. A hierarchical plan is a plan in which a top level node is expanded into sequences of lower level nodes each of which in turn may expand into yet lower level nodes. The leaves of the plan structure form a non-hierarchical plan of concrete actions the agent can actually take. Humans explicitly or implicitly maintain hierarchical plans all the time, such as when planning to buy milk, which expands into going to the store and purchasing milk, which in turn expands into walking to the car, getting in the car, driving to the store, and so on. Hierarchical plans have the advantage of making independence assumptions: if your goal is to buy milk, how you get to the store does not matter: you could walk, drive or bike.

Plans and planning are intimately related to perceived affordances. In fact, perceived affordances are the basis for planning. The current situation must contain an affordance predicting one could go buy milk, as otherwise one would not plan for it. Similarly, someone will only consider driving to the store, at a different level of affordances, if that person actually has access to a car, and if his or her encoding of the situation contains the perceived affordance of driving. Perceived affordances are thus not the elements of a plan, but at each step they are the possible choices a planner faces when making decisions. Thus a planner must maintain sets of perceived affordances to perform its planning, and a hierarchical planner maintains hierarchical trees of perceived affordances.

Parsing for Plan Recognition

Our implemented representation of perceived affordances is based on context free parsing. Context free parsing recovers hierarchical structures from a sequence of non-hierarchical observations, so it is natural that context free grammars, and especially PCFGs have been suggested as ideal paradigms for performing hierarchical plan recognition (Bobick and Ivanov, 1998; Pynadath and Wellman, 2000), a suggestion that originally dates back at least to Miller et al. (1960). In this case, the symbols in the terminal string correspond to observed events in a temporal sequence, and the grammar specifies possible higher level event structures. The simplified example depicted here is based on the studies that will be described in the next section. The example involves two players, Roirry (prefix 'R') and Isania (prefix 'I'), that engage in the short sequence of events depicted in Fig. 1. Isania pulls a lever to open a door, and Roirry goes through the door and fetches a key from a chest in the next room. A context free grammar parser produces the parse tree shown in the upper part of Figure 1 given the input symbol sequence depicted in that figure and the grammar in Table 1.

Table 1: Sample Plan Recognition Grammar Fragment

R_RETRIEVE_KEY → R_ROOM_1_TO_ROOM_2 R_OPEN_CHEST R_TAKE_KEY
R_RETRIEVE_KEY → R_ROOM_1_TO_ROOM_2 R_OPEN_CHEST R_TAKE_KEY
R_ROOM_1_TO_ROOM_2 → I_MAKE_DOOR_PASSABLE R_ROOMCHANGE_ROOM_1_TO_ROOM_2
R_ROOMCHANGE_ROOM_1_TO_ROOM_2 → R_THROUGH_DOOR R_ENTER_ROOM_2
I_MAKE_DOOR_PASSABLE → I_PULL_LEVER O_OPEN_DOOR
I_MAKE_DOOR_PASSABLE → I_BREAK_DOOR
I_MAKE_DOOR_PASSABLE → I_UNLOCK_DOOR I_OPEN_DOOR
R_OPEN_CHEST → R_UNLOCK_CHEST R_LIFT_LID
R_OPEN_CHEST → R_BREAK_CHEST

As we aim to use the internal states of a plan recognizer to represent a set of affordances, we need to be careful to select an algorithm that does predict all possible interactions at all levels at any given point in time, but that uses the symbols observed to constrain its search. The ideal candidate for an efficient parser along these lines is an Earley parser, which performs a combination of top-down prediction and bottom-up completion of parse trees to optimize its search behaviour (Earley, 1970). An Earley parser is based on the notion of an Earley state, a structure that concisely summarizes the state of the parser at a particular point in the observation sequence, and at one level of the current parse. That is, each state concerns exactly one grammatical rule, and marks how far in this rule the parse has progressed. In the probabilistic Earley parser we use here (Stolcke, 1995), a state also computes a probability of the grammar having produced the input sequence so far, and entering this state. We can take a state that has not completed its grammatical rule (i.e. that still has symbols left to parse in the tail of the rule) to be predicting the next symbol in the tail with a certain probability. These states are coloured green in Figure 1. Complete states (that have successfully advanced beyond all the symbols in the rule tail) correspond to the nodes in a parse tree - they assert that a higher level element of the hierarchy is indeed present in the input stream. These are coloured blue in Figure 1.

A parse state in an Earley parser used for plan recognition is an ideal candidate for a computational manifestation of a perceived affordance. Assuming that the parser is used to recognize the plans of a particular agent, it 1) predicts possible future interactions with the world at a particular point in time (the symbols to the right of the dot in the state), 2) ranks the likelihood of possible future interactions given the interaction seen so far through its forward probability, 3) applies to a particular level of abstraction, but is related to other levels due to the hierarchical nature of the grammar and 4) summarizes a segment of past interaction to predict the future. As an Earley parser progresses, it maintains complete state sets for each point in time, thus providing a complete history of past actions and predictions in addition to currently relevant predictions. We call the grammar used by this Earley parser an *affordance grammar*. This grammar is a predictive model of the structure of the world, representing a certain agent's predictions about and possible interactions with the world.

Language Grounding

The linguistic parsing step uses the same Earley parser as described earlier, this time parsing a string of words. Whenever the parser produces a complete state, it attempts to ground the constituent just produced in terms of ABCs. For this purpose, constituents can be associated with *concept definitions*. A concept definition takes the form of a nested function call expressing how the current set of perceived affordances is to be filtered to arrive at the ABC for this constituent. Every lexical item can be associated with one or more non-nested function calls. Such a call includes the name of the filtering function to apply to the set of affordances, and the argument positions used in the function call. Upon completion of a grammatical rule, the parser attempts to form a valid nested function call from the call present in the tail of the rule. Figure 2 shows a simple parse tree with associated filter functions. This method of incremental composition driven by language syntax is akin to other work that associates grammatical rules with lambda calculus expressions (Schuler, 2003) and our own work that performs compositional grounding according to explicit composition rules in the grammar (Gorniak and Roy, 2004).

An utterance occurs at a specific point in time, and at that time the plan recognition Earley parsers will have a particular set of current and past Earley states under consideration. To interpret a concept definition, the nested func-

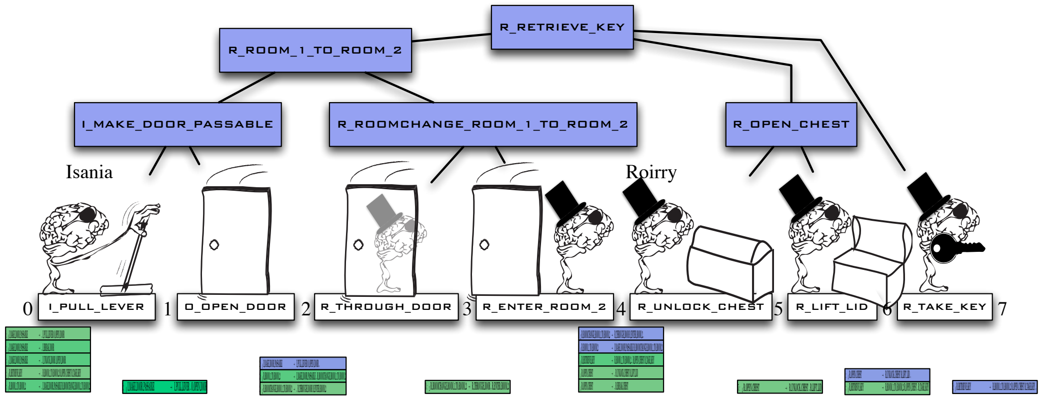

Figure 1: Sample Event Trace and Plan Parse Tree

tion call it represents is interpreted as an incremental filter on the full set of perceived affordances. Thus, for example, a noun like "gate" might select all interactions involving opening, unlocking, breaking and walking through doors at all present and past points in time, whereas a verb like "open" might filter these to only include the possible and actual interactions of opening doors. This simple example is shown in Fig. 2. Fig. 3, on the other hand, shows the filter expressions from this simple parse tree applied to the previous affordance example. In sequence, the selected affordances for $select(DOOR)$, $select(OPEN)$ and $select(OPEN, select(DOOR))$ are highlighted.

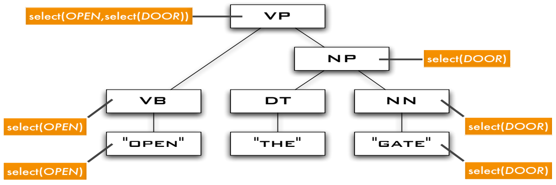

Figure 2: Simple parse tree example and affordance filters

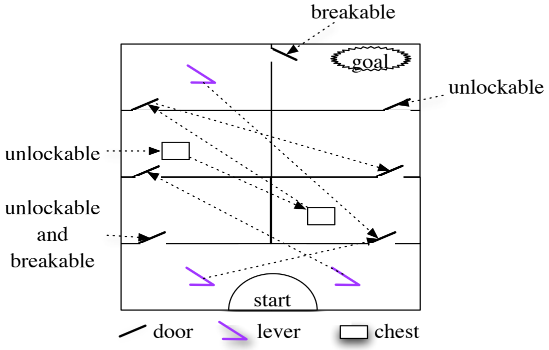

Figure 4: The map of the module used in studies.

Studies and Results

We have evaluated our implementation of the ABC theory by employing it to interpret situated language recorded from human-human communication during co-operative game play. To do so, it is not only necessary to record and analyse human language, but also to apply the machinery introduced in the last section to model the situation in which the language occurs. Here, we turn to multi-user graphical online role playing games to provide a rich and easily sensed world to support and capture human interaction.

We describe a set of studies using a commercial game, Neverwinter Nights, that includes an editor allowing the creation of custom game worlds. Fig. 4 shows the map used for the study presented here. Dependencies between objects in the map are indicated with dotted arrows (for example, an arrow links and lever that opens a door, and any chest containing a key for another chest or a door). The only objective of the puzzle is to reach the goal indicated on the map. When the two players start the puzzle, they only know that there is a goal they need to step on somewhere in the module. This puzzle is designed for players to separate and communicate their instructions and goals by using typed text. To limit dialogue phenomena, which are not the focus of the study, one of the players is randomly chosen in the beginning and forced to only use the following phrases: "Yes", "No", "I Can't", "Done", "Now", "What's going on?", "OK." The other player is free to use unrestricted language.

The study included 26 players who played in 13 dyads after responding to ads on the bulletin boards on the Neverwinter Nights website. Eleven of these dyads completed the puzzle in times ranging from 25 minutes to one hour, whereas the others gave up after one hour. Nine sessions served for development purposes, such as writing the affordance grammar and training the linguistic parser, and a group of four sessions formed an unbiased evaluation set. We first annotated the development data and built and trained the system, then annotated the evaluation data and tested the implementation on this previously unseen data.

Events presented to the plan recognizer were at the level of opening a door, walking into the next room or activating a lever. The recognizer used a grammar of about 6500 probabilistic rules that were automatically generated from a hand-coded grammar of about 90 meta-rules capturing the structure of the puzzle as shown in Figure 4. The current evaluation focuses on directives because their effect on the second human player is relatively easy to measure. We annotated 302 utterances in the development sessions and 69 utterances in the testing sessions as directives, and all results shown here are from these sets of utterances. The language parser uses 13 different affordance filter functions, including ones capturing

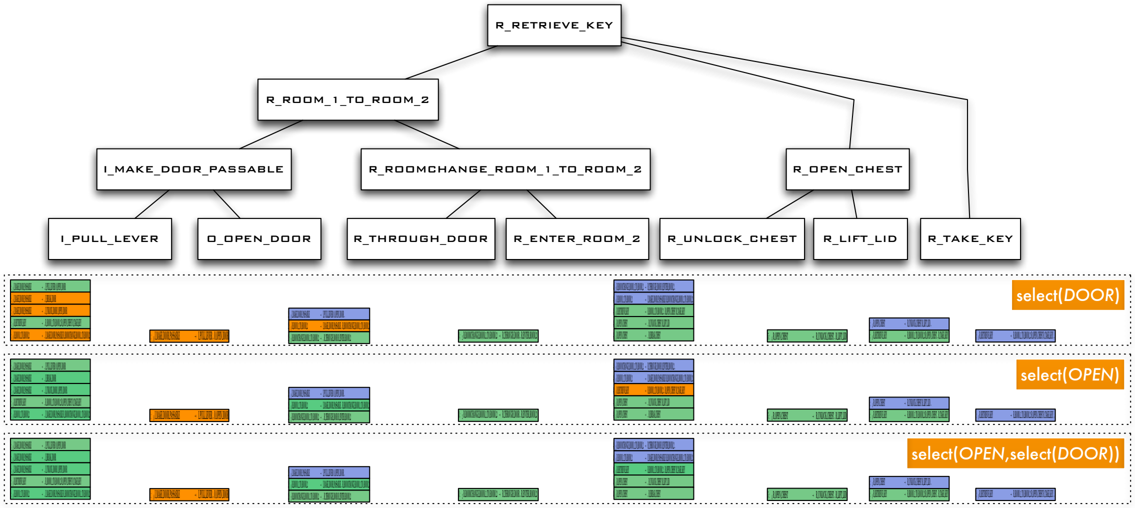

Figure 3: Filter functions applied to affordance example

spatial location ("west", "left"), spatial indexicality ("this", "that"), temporal indexicality ("let's try that again") and possession ("my lever") encoding some specific meanings for the game context. For example, a lever belongs to a player ("my lever") if the player was the one to most recently interact with it.

Whenever one player gives the other a directive, the utterance is parsed by the language parser to produce an affordance filter specification. The plan recognizer then runs this filter specification on the complete set of affordances produced up to this point in the game, which yields a filtered set of affordances. These are then ranked using their probabilities. If the best result is not currently applicable, the system plans towards a state in which it would be applicable and makes its prediction the first step along that plan. To measure performance, the final prediction is compared to the next action the player in question actually takes, and counted as correct if it matches.

The first row of Table 2 (*All Directives (AD)*) shows the performance on the complete set of directives. However, players do not always follow instructions, so the second row (*Followed Directives (FD)*) shows performance only on the 281 cases where the player actually performs an action that matches the directive as determined by the annotator (64 in the testing session). As many directives are short utterances marking the time to take and action rather than providing linguistic content (we call them *action markers*), the row *Followed Long Directives (FLD)* in Table 2 shows performance on the half of the directives that contain more than one word. The gap to the pure plan recognition baseline widens significantly on this utterance set, showing that the system can understand more complex language and produce the correct concept for many of these directives.

Table 3 shows a number of prediction baseline results for the same data sets. Row 1 of Table 3 shows the performance if language is ignored - that is, if we simply pick the most probable prediction of the plan recognizer at the point an utterance occurs, without paying attention to the words in the utterance. As above *Plan Recognition (FLD)* restricts the pure plan recognition baseline to those directives that were correctly acted upon by the listener and use more than one word. Finally, *State Based Random* randomly picks amongst all the actions players were ever observed to perform in a room combination.

Table 2: Accuracy Results for Understanding Directives

Selected Utterances	**Development**	**Test**
All Directives (AD)	70%	68%
Followed Directives (FD)	72%	70%
Followed Long Directives (FLD)	61%	68%

Table 3: Prediction Baselines

Prediction Type	**Development**	**Test**
Plan Recognition (AD)	65%	63%
Plan Recognition (FLD)	50%	60%
Random (AD)	15%	17%

When interpreting these results, it is important to keep in mind that perfect prediction cannot and should not be achieved in any of these cases. The puzzle naturally causes much exploration by the players, and, as will be discussed further below, situations and directives often do not limit players to a single next action. Some amount of variability is thus inherent in the scenario. The best overall performance of the complete system was 72%. Given the complexity of the problem and the leeway players appear to give each other in following their own utterance, this figure indicates that the theory and implementation presented in this paper make for an effective substrate for language understanding systems.

The low random baseline shows that prediction is no simple task (even this baseline does not pick amongst all possible actions, but only those players performed in the development data). Language understanding heavily relies on plan recognition - often the meaning of an utterance is highly constrained by the player's states and plans. Taking the words into account, however, improves again on the pure plan recognition performance. The best measure of this improvement is the 11% gain (8% in the test set) seen when considering the set of correctly followed directives longer than one word.

The performance gain is smaller when considering all utterances because performance is dominated by action markers, for which linguistic content plays little role, and thus yields no improvement in performance.

Performance on the test utterances is entirely comparable to that on the development utterances, showing that the plan recognition grammar and linguistic parser, while restricted in their coverage, generalize well to unseen data. Of note is that individual sessions differ greatly in playing and communication style. In fact, there is a single session in the test set that contains very repetitive and easily predicted player behaviour. When it is omitted, the test set performance baselines are equals to or lower than the development set baselines.

Conclusion

We have outlined a theory of concepts based on perceived affordances: structured units of mental representation that make predictions about possible interactions. We believe the ABC theory to be a useful new view of mental representation of concepts. It is unique in its computational interpretation of Gibsonian affordances based on plan recognition, and its successful realization in a language understanding task dealing with spontaneous, situated human language. The implementation presented in this article provides a convenient framework for probabilistic hierarchical reasoning about affordances while understanding situated language. It will be important to integrate this framework with other approaches and views on affordances (Steedman, 2002; Roy, 2005) and to re-phrase existing approaches dealing with other aspects of grounded language understanding in an affordance-based framework.

References

Bobick, A. F. and Ivanov, Y. A. (1998). Action recognition using probabilistic parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Carey, S. (1985). *Conceptual Change in Childhood*. MIT Press.

Carnap, R. (1932). Überwindung der metaphysik durch logische analyse der sprache. *Erkenntnis*, 2.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, 6(8):451–455.

Fleischman, M. and Roy, D. (2005). Why are verbs harder to learn than nouns? In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

Gibson, J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., editors, *Perceiving, Acting and Knowing*, pages 67–82. Wiley, New York.

Gorniak, P. (2005). *The Affordance-Based Concept*. PhD thesis, Massachusetts Institute of Technology.

Gorniak, P. and Roy, D. (2005a). Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the International Conference on Multimodal Interfaces*.

Gorniak, P. and Roy, D. (2005b). Speaking with your sidekick: Understanding situated speech in computer role playing games. In *Proceedings of Artificial Intelligence and Digital Entertainment*.

Gorniak, P. and Roy, D. (in review, 2006). Situated language understanding as filtering perceived affordances. *in review*.

Gorniak, P. J. and Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. In Margolis, E. and Laurence, S., editors, *Concepts: Core Readings*, chapter 1, pages 3–81. MIT Press.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Adams, Bannister, Cox, New York.

Narayanan, S. (1997). *KARMA: Knowledge-based Action Representations for Metaphor and Aspect*. PhD thesis, University of California, Berkeley.

Plunkett, K., Sinha, C., Moller, M., and Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4(3&4):293–312.

Prinz, J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. MIT Press, Cambridge, MA, USA.

Pynadath, D. V. and Wellman, M. P. (2000). Probabilistic state-dependent grammars for plan recognition. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI2000*. Morgan Kaufmann Publishers.

Regier, T. (1996). *The Human Semantic Potential*. MIT Press.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104:192–233.

Roy, D. (2002). Learning words and syntax for a visual description task. *Computer Speech and Language*, 16:353–385.

Roy, D. (2003). Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*.

Roy, D., Gorniak, P. J., Mukherjee, N., and Juster, J. (2002). A trainable spoken language understanding system. In *Proceedings of the International Conference of Spoken Language Processing*.

Schuler, W. (2003). Using model-theoretic semantic interpretation to guide statistical parsing and word recognition in a spoken language interface. In *Proceedings of the Association for Computational Linguistics*.

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

Steedman, M. (2002). Formalizing affordance. In *roceedings of the 24th Annual Meeting of the Cognitive Science Society*, pages 834–839.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Stoytchev, A. (2005). Behavior-grounded representation of tool affordances. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, page ??

Yu, C., Ballard, D., and Aslin, R. (2003). The role of embodied intention in early lexical acquisition. In *Proceedings of the Cognitive Science Society*.