Learning and Playing in Wubble World

Wesley Kerr and Paul Cohen and Yu-Han Chang
USC Information Sciences Institute
4676 Admiralty Way
Marina Del Rey, CA 90292

Abstract

Children do not learn the meanings of words from parsing and understanding gigabytes of text; instead meanings are learned from competent speakers who relate language to what's happening in the child's environment. We present a word learning algorithm that operates in a video game environment where the players fill the role of the competent speakers and train softbots to learn language as a child would. We provide empirical evidence that the word learning algorithm successfully learns the meanings for some words in this environment and the children enjoy playing the game.

Introduction

How do children learn word meanings so quickly? Paul Bloom, a psychologist who has made a thorough study of the question, answers this way:

> *Young children can parse adult speech (or sign) into distinct words. They think of the world as containing entities, properties, events, and processes; most important they see the world as containing objects. They know enough about the minds of others to figure out what they are intending to refer to when they use words. They can generalize; and so when they learn that an object is called "bottle" and an action is called "drinking", they can extend the words to different objects and actions. They can also make sense of pronouns and proper names, which refer to distinct individuals, not to categories; and so they understand that "Fido" refers to a particular dog, not to dogs in general. (Bloom 2001)*

The purpose of the Wubble World project is to have a computer learn language as young children do. Children learn language from competent, facilitative speakers who strive to associate language with what's going on in the child's environment. Some aspects of the semantics of words and phrases are immediately accessible when the language refers to the scene. For a computer to learn language as a child does, it needs perceptual access to an environment about which a competent language-user is speaking (or typing), and it needs learning algorithms that can associate utterances with what's going on in the environment. This will not be enough, according to Bloom, because children also associate words with what they infer to be happening in the mind of the speaker. However, this paper shows that some word meanings can be learned associatively given a parallel *sentence-scene corpus* containing sentences uttered about the environment, and the states of the environment when the sentences are uttered.

The common way to acquire a sentence-scene corpus is to generate many instances of activities with robots or simulated agents and describe each with sentences in natural language (Oates 2003; Siskind 2001; Regier and Carlson 2001). This approach requires a serious investment of time and in some cases trained individuals to annotate activities.

According to a study by the NPD group, 63% of the US population plays video games (NPD 2007). Some researchers speculate that the collective efforts of these players might produce vast amounts of data to solve various AI problems (Buro and Furtak 2005; Dunham, Forbus, and Usher 2005). Our approach to building the sentence-scene corpus is to provide an enjoyable game in which children interact with softbots in natural language. The children are told to treat their softbots —- called wubbles — as younger siblings who might need help in understanding what is said to them. As the children type instructions to their wubbles, we record the states of the Wubble World environment. This produces a parallel corpus of sentences and scenes.

In (Kerr et al. 2007), we demonstrated that wubbles could learn meanings for nouns, adjectives, and prepositions, but the sentences were generated automatically, not by children playing Wubble World. Here we present the results of having children train their wubbles using natural language.

Previous Work

Related work begins with SHRDLU, a blocks world created by Terry Winograd (Winograd 1972). This system generates and understands natural language situated in a simulated world. SHRDLU acquired definitions of new words through its interactions with people, but all interactions lacked a gameplay component. Rather than operating directly on the 3D representation of the simulation, reasoning was carried out using a symbolic abstraction. Word learning in Wubble World occurs in a similar fashion.

Gorniak and Roy (Gorniak and Roy 2007) used the Neverwinter Nights engine to generate a puzzle game played by

two people. Only one of the players was allowed to talk, and all communication between the players was recorded. The authors developed a software system that is able to understand language after gathering data from several subjects solving the same task. This system can interface with the original puzzle game and through voice communication with the first player execute commands as the second player would in the game. This approach of gathering data and refining semantics is similar to ours, although we use an on-line model of word learning that allows instruction from a teacher.

In a similar effort, Orkin and Roy (Orkin and Roy 2007) developed an on-line social game between two people. The interactions of the players took place in a virtual restaurant. The authors gathered linguistic data from the players' typed messages, as well as scene descriptions such as character positions, the actions the characters performed, and so on. The authors refined the gathered data into a compact representation, called Plan Networks, that can accurately judge typical behavior versus atypical behavior within the virtual restaurant. The research generates large amounts of data, but there has not been a signficant investment in learning the meanings of words used within the environment.

Sankar and Gorin (Sankar and Gorin 1993) developed a 2D environment similar to Wubble World. It displayed a set of objects in a scene, and provided an interface for interacting with a softbot represented onscreen as an eye. The eye was directed towards objects in the scene using natural language. The system successfully learned 431 words from over 1000 conversations, on-line, by adjusting a connectionist network. Other researchers worked with the connectionist paradigm to explore the symbol grounding problem (Schyns 1991; Plunkett et al. 1992; Gasser and Smith 1998; Regier 1996). Most of this research resulted in networks that were trained and tested, in batch, to associate words with images. In this paper we present a different word learning algorithm that learns on-line and also operates in a 3D environment.

Problem Statement

We set out to design a game entertaining enough for children to play for its own sake, and that would also generate a sentence-scene corpus. The result is Wubble World, an on-line game environment where children play the role of parents and softbots called wubbles play the role of infants. Wubble World consists of several mini-game environments, but this paper focuses on just one of them, the Wubble Room mini-game, in which children direct their wubbles in natural language to perform tasks. Before playing the game children are told: "This is your wubble. It knows some English, but not much. Often, when you tell it to do something, it will ask you what you mean. You will be given a task for your wubble to do, and you will have to tell it in English how to do the task."

The Wubble Room mini-game has three separate rooms: The *Introduction Room* is where children familiarize their wubbles with different objects. The children can interact with their wubbles and four other objects in the room. After a short period of interaction in the Introduction Room, a larger set of varied objects is placed in the room. The wubble begins questioning the child about these objects, and in the process gathers additional data with which to refine recently learned concepts, such as "red" and "cube." The wubble gathers this information with assertions like: *Click on all of the red objects*.

After this, the wubble and child are promoted to the *Apple Room*, which contains several objects differing in size, shape, and color. On a platform in one corner of the room is an apple. The child's goal is to get her wubble to retrieve the apple. The physics of Wubble World won't let wubbles jump as high as the platform, so the child must instruct her wubble to build a staircase from the objects in the room.

The final room is the *Sorting Room* where the child works with her wubble to line up the objects in the room from tallest to shortest.

Children instruct their wubbles in English to perform tasks. Through this interaction, the wubbles learn nouns (*box*, *cone*), adjectives (*red*, *tall*), and spatial prepositions that describe relations between objects (*behind*, *left of*). All wubbles are provided with the meanings of a small set of verbs *a priori*. So if a child says, "pick up the red cube," the wubble knows what "pick up" means and can perform the action.

Word Learning

Wubble word learning is illustrated in Figure 1. Word learning is perhaps a misleading phrase because what's learned is a concept, the name of which is a word or a phrase. Wubbles maintain simple, feature-based representations of concepts. For objects, the features are shape, color and size; for prepositions, the features describe ways of dividing up space. Each feature has an associated weight or probability distribution, as shown in step 2 of Figure 1. For instance, the weights associated with the feature values $shape = cylinder$ and $color = blue$ are higher than the weights associated with $shape = sphere$ and $color = orange$.

The wubble learns incrementally, on-line, by updating the weights associated with each feature. When a wubble encounters a sentence, it parses it (step 1, Fig. 1) and retrieves from memory the concepts that correspond to the words and phrases in the sentence (step 2, Fig. 1). If a word is new, the wubble creates a set of weight distributions, one for each feature, initialized with a uniform distribution. Next, the wubble *imagines* an object (or adjective, or preposition) by sampling from the retrieved distributions of each feature. Sampling is probabilistic, so there is a small probability that the wubble will imagine a blue cone, say, instead of a blue cylinder (the output of step 2, Fig. 1). It then compares the imagined object with the objects in the scene, and it calculates scores for the correspondences between the imagined object and those in the scene (step 3, Fig. 1). After that, if the wubble is certain enough it has identified the referent of the word or phrase, it acts. For instance, in the scene in Figure 1, the wubble would go to the blue cylinder. However, if it remains uncertain about the referent of a word or phrase, it asks the player to point to the referent (step 4, Fig. 1). Then the wubble updates the feature representation associated with the word or phrase to make it more like the

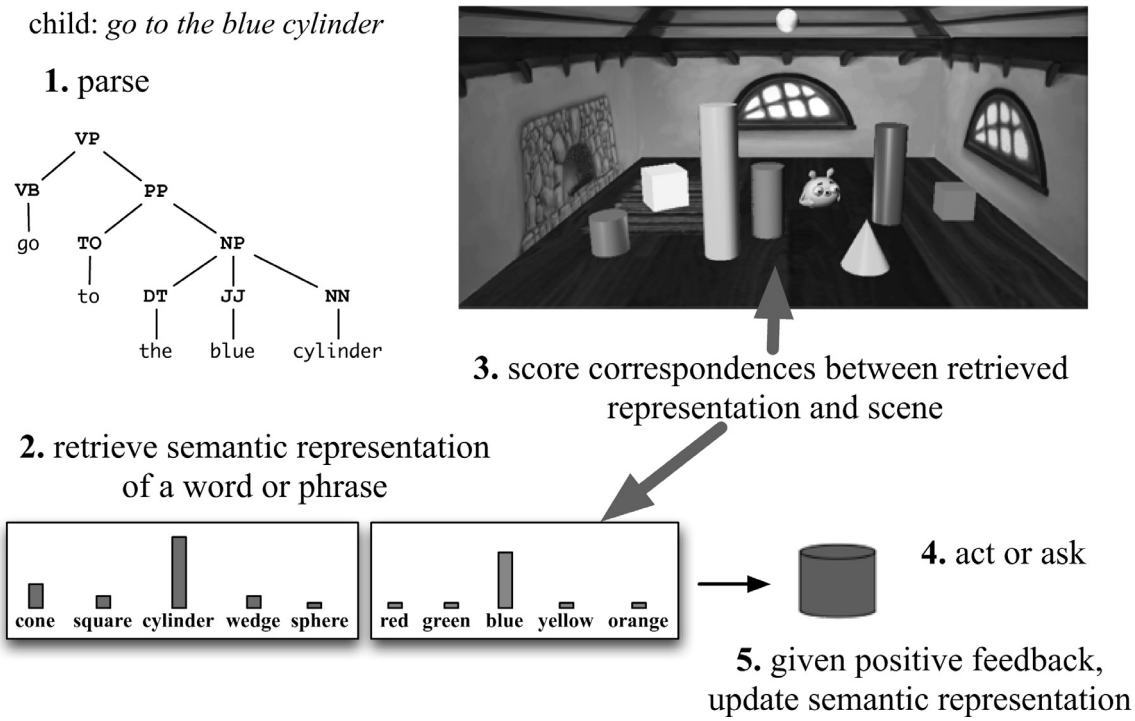

Figure 1: The steps involved in learning word meanings in Wubble Room.

object in the scene (step 5, Fig 1). In this way, it quickly learns the feature representations that should be associated with words. A more formal description of word learning in Wubble World may be found in (Kerr et al. 2007).

Experimental Design and Protocol

We designed an experiment to test two related hypotheses. The first is that while a child is playing the game, the training information she provides will help the wubble learn the correct meanings of words. The second hypothesis is that children will enjoy playing the game.

We hired 10 children between the ages of 11 and 14 to participate in the experiment. Each was paired with an experiment proctor who introduced the child to Wubble World tasks and answered any questions. All of the children's interactions with Wubble World were confined to the Wubble Room area described earlier. Interactions were in English, which the children typed without difficulty on a standard computer keyboard. There was only one restriction on the children's utterances: They were told that the wubble could only understand half a dozen verbs: turn, walk, jump, pick up, put down, and choose. The wubble would not do anything if the children used other verbs. This ensured that the wubble could act on all of the children's utterances.

As described earlier, each child started with a wubble in the Introduction Room, then moved on to the Apple and Sorting rooms. We did not restrict the length of the experiment, and several children spent more than two hours working with their wubbles in the Apple and Sorting rooms. Although the tasks to be done in these rooms did not change, the wubbles' ability to understand English improved, and this kept some children engaged.

To get enough interactions between children and wubbles we merged all students' sessions together — each session being a sequence of sentences typed by the child to her wubble. For each sentence, the word learning algorithm worked on learning the meaning of each constituent noun, adjective, and preposition. A word and its associated scene constitute a single training instance.

We selected 23 words from the children's transcripts and specified an "ideal" concept for each. Each concept is described by a set of features. A feature is *defining* iff every instance of the concept must have the same value of that feature. For example, every blue object must have the value "blue" for its color feature. An ideal feature vector representation of a concept is constructed by setting the probability of each defining feature value to 1.0, and setting the probability distributions over values of non-defining features to reflect the unconditional probabilities of these feature values in the environment. For example, if there are only ten cubes and five cylinders in the environment, then the ideal feature vector representation for blue objects would have $Pr(color = blue) = 1$, and $Pr(shape = cylinder) = .333$, $Pr(shape = cube) = .666$.

To show that the wubbles learned the concepts associated with words, we compared the learned concepts for the 23 words to the ideal feature-based representations using the Kullback-Leibler (KL) distance (Kullback and Leibler 1951). Smaller KL distances indicate that the feature representation of a concept is more similar to the ideal feature

representation.

After the experiment we debriefed the children to get a qualitative analysis of the game — what the children liked, what worked and didn't work, how the children would improve the game, and so on.

Experiment Results

Recall that we tested whether it is possible to learn approximations to the ideal feature based representations of 23 words. Aggregating the training instances from all students yielded 957 sentences in which one or more of these words appeared. Figure 2, shows the *average* KL distance, over the 23 words, after successive training instances. Although each training instance affects the learned representation of just one word, we see that the average KL distance decreases with training, indicating that, for these 23 words, their feature based representations approach the ideal representations.

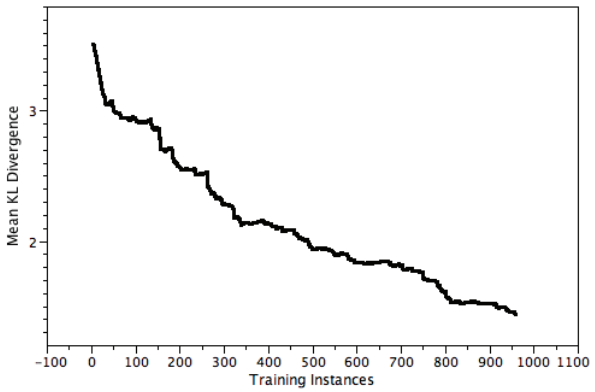

Figure 2: Average KL distance between learned word meanings and ideal meanings for 23 words, given children's interactions in English with their wubbles.

Not all word feature vectors followed an inexorable trajectory toward the ideal feature vector. Tracking the KL distance for the word "blue" we observe in Figure 3 that divergence decreases rapidly, then starts to grow, slowly. This is because the final eight training instances of a "blue" object were all the same size. The word learning algorithm learned that a "blue" object is not only *blue* but also a particular sized cube. More training instances, in more variable environments, will fix this problem.

Of course, the children used many more than 23 words in their interactions with their wubbles. We did not define ideal feature based representations for all of these words, but we did attempt to characterize whether the word learning algorithm was learning *some* definition for these words, even if it wasn't what we considered the correct definition. One such assessment is given by the entropy of the feature vector representation of a word. High entropy means a uniform distribution of mass over all the features, or high uncertainty about which features define a word. So if the entropy for a word decreases as the number of training instances increases, it means that the wubble is becoming more certain

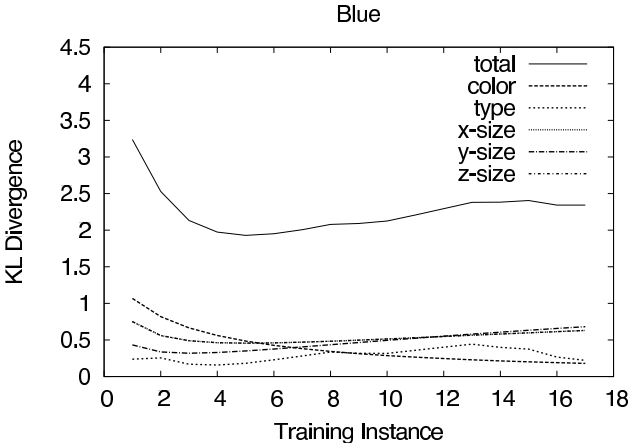

Figure 3: KL distance of "blue" over 17 training instances.

that particular features define the word.

To illustrate, consider the word "end," used by just a handful of children. Wubbles cannot learn the common meaning of "end" because none of the word features describes the relationships between objects and regions of space that are denoted by the word.

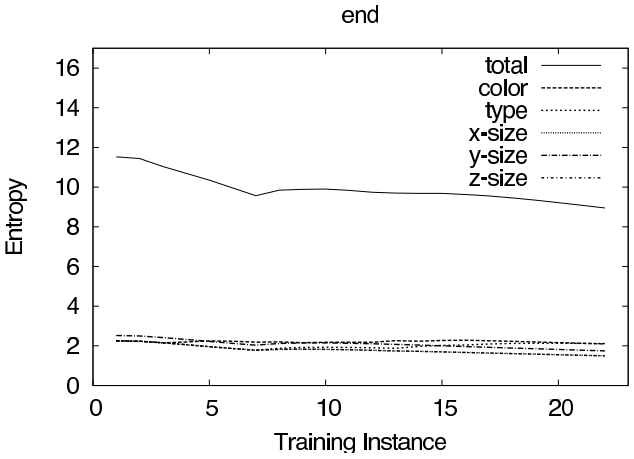

Figure 4: Final entropies of features of "end" after 23 sentences.

One girl persisted in trying to teach her wubble "end" with sentences like, "move to the end of the orange cubes" (denoting a location at the end of a row of cubes). Figure 4 shows the entropy of each of five word features – color, size-in-x, size-in-y, size-in-z, and geometry – over 23 sentences (end-size-x and end-size-y have identical entropies throughout). The total entropy, which is just the sum of the entropies of the five features, decreases, although there is a bump, caused by increasing entropy of the geometry feature, in the first ten sentences. It appears that "end" is becoming associated with a particular object (an orange one) at the end of a row of objects. The interesting thing about this is that the speaker used the word "end" to refer to two locations in Wubble Room, one near the end of a clump of cylinders, the

other near the end of a row of cubes. In both cases, the nearest object to the end was short and orange, and this is the meaning that the wubble learns for "end." Figure 5 shows the final values for the five word features. Orange is the dominant color, geometry is pretty uniform, and size-in-y is distinctly "small."

We discuss the limitations of wubbles' word features in the conclusion, yet despite these limitations, children had the impression that their wubbles were learning word meanings. In the previous example, the girl who trained her wubble to understand "end" thought that the training was successful. It was not until later, when we debriefed her, that she came to understand that the wubble had associated "end" with particular objects. She apparently *wanted to believe* that her wubble had learned the common meaning of "end." This illusion has been documented in other contexts, as well, notably Dennett's article about his experiences judging the Loebner Prize competition (Dennett 2004). Apparently, we cannot help ascribing intelligence to and intentionality to artifacts.

This probably explains why Wubble Room, which is really quite dull as computer games go, captivated seven of the ten children who played it. (It was difficult to get reasons from the other three, but our impression is that they just did not think the game was interesting or engaging.) Those who did all mentioned how cool it was to *teach* the wubbles, and they were impressed that the wubbles learned. As demonstrated by Wubble Room, the *Black & White* series, and the *Creatures* series, we suspect that having game characters learn, especially with players teaching them, will continue to be an opportunity for game developers.

Conclusions

Although the learning algorithm is in some ways successful — it quickly finds correct representations of some nouns, adjectives and prepositions — it cannot learn other words because the underlying feature vector representation is poor. In particular, the representation cannot capture relations between objects (e.g., "the *bigger* one") or composite objects (e.g., "the *stack* of blocks"). It cannot handle pronouns or many other words classes; although we are currently extending the representation to handle verbs. The child's theory of mind — the ablility to infer what the speaker is thinking about — is essential to Bloom's theory of word learning, but Wubbles have no theory of mind, and cannot understand words like "want" in sentences like "the one I *want*."

These are limitations of the underlying representation and not necessarily limitations of the learning algorithm. In fact, it appears that the main steps of the algorithm — retrieving a word representation from memory, imagining an object or relation by sampling from that representation, comparing what's imagined with objects and relations in the scene, and then updating the word representation to make it more like what's found in the scene — allows wubbles to learn word meanings from positive instances quite quickly. Moreover, we demonstrated that children could teach their wubbles and enjoyed it.

We built a more complex environment that afforded a wider range of activities and included two agents, a wubble controlled by the child and a dragon that appeared to the child to be autonomous but was in fact controlled by a graduate student in another room. The wubble and the dragon had different abilities and had to cooperate. We learned that more adept agents in a more complex environment evokes a bigger vocabulary and more semantic distinctions. New verbs, such as "open," appeared, and we saw verbs paired with modifiers such as "jump on" versus "jump over." The children began using verbs whose semantics are plan-like, such as "fetch," which is composed of "go to", "pick up", and "put it here" actions.

This Wizard-of-Oz experiment also supported Bloom's hypothesis that word learning is scaffolded by a theory of mind. In one version of the experiment we replaced the words typed by the children with nonsense words. The graduate student who received these nonsense sentences was able to guess at their meanings simpy because his dragon and the child's wubble were working cooperatively on a task.

Development is underway on a better semantic representation of words. Ideally the representation will be robust enough to allow the addition of new semantic features. As the world becomes more complex, the semantics captured by the representation should automatically become more complex. A separate effort to generate better semantic representations is also underway, but focuses on semantics in isolation, independent of the environment, e.g. Proposition Bank (Palmer, Gildea, and Kingsbury 2005), OntoNotes (Pradhan et al. 2007), and Cyc (Matuszek et al. 2006).

Finally, we need to build better games. One example is a Wubble World multiplayer on-line game called "Sheep." Children are divided into two teams and interact within each team through voice chat. The communication is logged on our central servers as the team members work together to herd sheep into pens, and try to rustle sheep from the other team. These captured interactions are processed offline to learn the meanings of words. The game is popular, but there are some kinks with the speech-to-text software, which doesn't work well with eleven-year-old girls chirping about wubbles. The sentence, "Hey Lauren, Hey Lauren, look there, get that one, you know like the one nearest you" translates to, "Shell Oil found new oysters in Azerbaijan." Nevertheless, we are hopeful that children playing online games will prove to be a great source of data for a variety of AI learning problems.

References

Bloom, P. 2001. Précis of how children learn the meanings of words. *Behavioral and Brain Sciences* 24:1095–1103.

Buro, M., and Furtak, T. 2005. On the development of a free rts game engine. *GameOn'NA Conference*.

Dennett, D. 2004. Can machines think? *Alan Turing: Life and Legacy of a Great Thinker*.

Dunham, G.; Forbus, K.; and Usher, J. 2005. nuwar: A prototype sketch-based strategy game. *Proceedings of First Artificial Intelligence and Interactive Digital Entertainment Conference* 45–50.

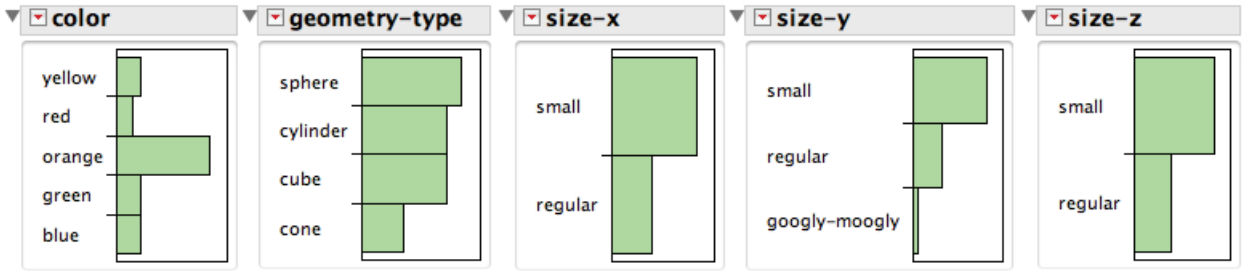

Figure 5: Final training distribution of "end" after 23 sentences.

Gasser, M., and Smith, L. 1998. Learning noun and adjective meanings: a connectionist account. *Language and Cognitive Processes* 13(2):269–306.

Gorniak, P., and Roy, D. 2007. Situated language understanding as filtering perceived affordances. *Cognitive Science* 31(2):197–231.

Kerr, W.; Hoversten, S.; Hewlett, D.; Cohen, P.; and Chang, Y. 2007. Learning in wubble world. *Development and Learning*.

Kullback, S., and Leibler, R. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*.

Matuszek, C.; Cabral, J.; Witbrock, M.; and DeOliveira, J. 2006. An introduction to the syntax and content of cyc. *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*.

NPD. 2007. Expanding the games market. `http://www.npd.com/press/releases/press_071212.html`.

Oates, T. 2003. Grounding word meanings in sensor data: Dealing with referential uncertainty. *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*.

Orkin, J., and Roy, D. 2007. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development* 3(1):39–60.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*.

Plunkett, K.; Sinha, C.; Møller, M.; and Strandsby, O. 1992. Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*.

Pradhan, S.; Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; and Weischedel, R. 2007. Ontonotes: A unified relational semantic representation. *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC-07)*.

Regier, T., and Carlson, L. 2001. Grounding spatial language in perception: an empirical and computational investigation. *J Exp Psychol Gen*.

Regier, T. 1996. The human semantic potential: Spatial language and constrained connectionism.

Sankar, A., and Gorin, A. 1993. Visual focus of attention in adaptive language acquisition. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93* 1:621–624.

Schyns, P. 1991. A modular neural network model of concept acquisition. *Cognitive Science* 15(4):461–508.

Siskind, J. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*.

Winograd, T. 1972. Understanding natural language.