

# Wubble World

**Daniel Hewlett and Shane Hoversten and Wesley Kerr  
Paul Cohen and Yu-Han Chang**

Information Sciences Institute  
University of Southern California  
Marina del Rey, CA 90292 USA

## Abstract

We introduce Wubble World, a virtual environment for learning situated language. In Wubble World children create avatars, called “wubbles,” which can interact with other children’s avatars through free-form spontaneous play or structured language games. Wubbles can also learn language from direct interaction with children, since the system uses principles from developmental psychology to restrict the complexity of this learning task: a shared attention model that includes deictic pointing, and a concept acquisition system that allows for rapid learning of new words from a limited number of exposures.

Since we have complete knowledge of the state and structure of the virtual environment, we are able to track correspondences between utterances and the scene in which they are uttered. This sentence/scene corpus will be a valuable resource as we attempt to tackle more sophisticated language learning tasks, such as the acquisition of syntax and verb semantics from world dynamics.

## Situated Language

Communicating with machines in natural language is hard. There’s been a lot of progress in recent years, mostly due to the success of statistical methods trained on huge corpora, but even with the fruits of these labors we still can’t communicate with machines in an intuitive way.

Why is it so hard for a machine to learn language? Part of the problem might be that there’s not enough of the right kind of data to feed to the statistical methods. Part of the problem might

be shortcomings in the statistical methods themselves, particularly in their focus on crunching language data without considering the physical context that elicits that data.

Take, for instance, the sentence “get out.” It’s clear that “get out” means different things depending on the context in which it is uttered. Indeed, the number of possible interpretations are staggering. By divorcing the utterance from the environment we lose the distinction between any of the flavors of “get out.” It seems unlikely that anyone could make much progress toward language understanding without first situating the learning process in a concrete environment. (See (Roy & Reiter 2005) for summary.)

Gathering data from these sorts of situated environments has been a daunting prospect up till now; the only option was to gather together a group of people, and put them in a monitored room to record and transcribe their interactions. This was both logistically daunting and expensive, and accounts for the paucity of large quantities of situated language data. These days, the popularity of high-speed internet and online games has made feasible a new approach to the collection of language data: if we can design the right sort of game, we can get people to provide us with the data we need, in an environment we could precisely describe and automatically record.

## Child’s Play

To this end we have designed an environment called “Wubble World” where children can play games and, in the course of playing, generate situated language data.

Using children as instructors offers several advantages: children use simpler language, which is easier for the virtual agent to learn. Children are likely to be comfortable interacting with virtual agents, and may also be willing to spend significant time engaging and teaching it. It is not unlike a child taking care of her doll; only this doll is interactive: you can talk to it, and it will respond.

Using children as models for language acquisition offers advantages of its own. Young children are not required to attain complete linguistic competence all at once; language capabilities are primitive at first and improve over a number of years. They do, however, accumulate steadily, and more sophisticated constructs (metaphor) are built on earlier foundations (spatial reasoning) (Lakoff & Johnson 1980). Wubble World provides this path for the learning wubbles by offering tasks and mini-games of increasing difficulty, which require the virtual agent to understand more complex commands as the interactions proceed.

### **The Importance of Being Social**

In humans, the most important mechanisms for language acquisition are social interactions between the child and the world, particularly those between the child and its mother. These interactions give a child the tools to manage the complexities of language acquisition. One example of such a tool is shared attention. Babies will automatically look where another person is looking; this helps the child resolve semantic ambiguities, as described above. For instance, a parent might say to a child: “Get me the *gartbalve*.” The child might not know what a *gartbalve* is, but could identify the object by following the parent’s gaze, or at least constrain its possible location to a restricted domain. The disambiguation can be strengthened by deictic pointing: gesturing to the *gartbalve* can define it for the child even more precisely. (See (Bloom 2001) for a summary of these results)

It’s clear that social learning has a lot to offer, and we have integrated some of these principles into our wubbles. Instead of dropping our agent into the world and leaving it to fend for itself, we let it learn by interacting with others. We give it a teacher, a (hopefully) patient teacher, who will work with it at the level of its competence, like a parent does with its child. Our agent can learn the names of objects in its environment, it can learn

names for important spatial relationships, it can learn simple concepts. This sort of learning is not sufficient for more ambitious language tasks, but more ambitious language tasks will certainly require this foundation. Speaking more generally, with enough language data collected from enough separate learning instances we’ll have a resource in situated language use that can motivate more sophisticated future efforts.

### **Wubble World**

We have designed a virtual world where large numbers of children can interact, called Wubble World. The child’s agent in Wubble World is a virtual creature called a “wubble.” There is precedence for this sort of virtual interaction: communities like Club Penguin ([www.clubpenguin.com](http://www.clubpenguin.com)) and Neopets ([www.neopets.com](http://www.neopets.com)) have proven to be very sticky, without offering much in the way of sophisticated functionality. Instead, they provide cute avatars the children can customize, some simple games, and the ability to socialize via a form of instant messaging. Wubble World employs these same ideas in service of a more principled learning goal.

There are currently two broad areas of Wubble World, each suited to a particular type of game. One of these is a language-learning room. In this room the child interacts with her wubble one on one, acting as its instructor (See Figure 2). Through these interactions the child teaches her wubble words for simple concepts: nouns, adjectives, and prepositions. The wubble needs to understand this language in the context of its environment in order to accomplish certain tasks; for example, it builds a stairway to reach a hanging piece of fruit that is otherwise out of reach. (Another system that uses natural language to identify object relationships is given by (Gorniak & Roy 2004))

The other game takes place in a virtual environment in which the wubble acts as the child’s avatar. The child, as her wubble, can explore the world and talk to other children, using an instant-messaging interface, about whatever she feels like talking about, or play a game, called Wictionary, that is itself situated in the virtual world (See Figure 1). In this area the interactions between wubbles are largely unstructured - the children have no externally-imposed goal to pursue - but we are bet-

ting that their natural social inclination will generate a large corpus of language data. This data can be analyzed with respect to the shared context provided by the details of the environment, and the wubbles' configurations within it. These games are described in more detail in the following sections.

### **Wictionary: Building the Sentence/Scene Corpus**

In the middle of Wubble World is a giant screen, much like the screen of a drive-in theater. This is the "stage" for the Wictionary game. Children sign up to play the game, where they take turns as *builders* and *guessers*. The builder's job is to specify a model and build it; the other players try to guess what is being built. When one of the players guesses correctly she acquires points, which can be spent to enhance the wubble avatar with new items and accessories, image customization, special powers, etc.

The models built in Wictionary are composed of simple geometric shapes, which the builder can rotate, scale, and color. These shapes can be combined into complex images, which appear on the screen as they are being manipulated; the effect is akin to watching someone build models out of construction paper.

Since the exact structural relationship of the objects in the Wictionary scene are known, and since the language used to describe them is recorded, we are able to compile a body of data relating one to the other: this is the sentence/scene corpus. At the moment most of the utterances are brief, aside from meta-game commentary (commentary on the images; jokes; general chatter) and tend toward simple concept names: "tower" or "cow."

We are currently working on adding dynamics to the Wictionary game, allowing children to move shapes around on the screen and thus act out scenes over time. We hope this will elicit richer sentences with more complicated syntax, and provide us with a semantic representation of dynamic language.

### **Blocks World Revisited**

On the surface, the learning room is reminiscent of "blocks world," a staple of traditional AI. In most formulations, it consists of several blocks of various shapes, sizes, and colors, whose state can be

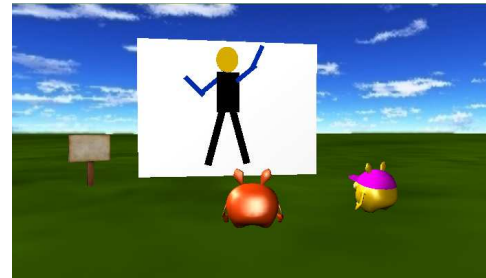


Figure 1: Wubbles playing Wictionary

described as a collection of predicates. The exercise is usually to construct a plan to transform the blocks from one configuration to another ("Put the green block on the yellow block"), or to answer queries ("Is the red block on the blue block?") using either natural language (Winograd 1972; Gorniak & Roy 2004) or logical propositions.

Blocks Room in Wubble World is slightly different. In this room, the Wubble represents a separate entity from the child, acting perhaps as a younger sibling. The goal is to teach language to the nascent wubble by describing the room and the actions that can be taken there. The wubble appears in the room along with some configuration of geometric shapes, which differs according to the learning task. In the most basic task there is a variety of different shape/size/color configurations, and the child must teach the Wubble to distinguish them, and also teach it prepositions, using natural language. Interaction with the wubble is required because it is the wubble, not the child, who can physically interact with the objects in the room. This division of labor is crucial to the elicitation of simple language concerning the scene from the child. For instance, the child might type: "Go to the right of the green cylinder." Initially the wubble doesn't know what *green* is, what a *cylinder* is, or where *to the right of* is. If the wubble doesn't know what to do it asks for help. ("Is this a cylinder?" "Is this green?" "Am I to the right of the cylinder?") The child can give the wubble feedback, either by typing a response or by indicating, using the mouse, which is the object in question.

The premise of this learning model is that it's OK to be confused; help is available. Just as children aren't expected to deduce the names of things without asking questions, the Wubble will need

help at first.

And just as children, once they possess a bit of language, can figure out the meanings of new words without explicitly being told, so the wubble can figure out which object is the “green cylinder” if it knows which color is green. The next section describes our implementation of an approach that allows the wubble to quickly and robustly learn concepts such as objects, prepositions, and procedures.

## Concept Acquisition

Like a child, a wubble forms concepts based on sensory experience. Sensory processing is a daunting problem in the real world, but as the wubble lives in a virtual world we have the ability to structure its sensory data so as to simplify the problem; consequently, every object it “sees” is a vector of the attributes COLOR, SIZE, and TYPE.

By discretizing these attribute values, the wubble can record its certainty that a particular word corresponds to a particular set of attribute values, e.g. (red, small, cube) or (red, medium, cone). For a given word, we formalize this by maintaining a vector of weights for each attribute,  $(W_i^a, \dots, W_n^a)$ , where  $i \in I^a$  are indices corresponding to each discrete value of attribute  $a$ . The individual weights  $w_i^a$  are updated to reflect the likelihood that this particular attribute value is observed given an uttered word. When given a word, the wubble chooses an object partly based on the probability distribution defined by the weights, as will be described later. The updating is performed so as to minimize the *regret* of the wubble as it chooses objects to associate with the uttered words, i.e. to minimize the loss of utility associated with choosing an incorrect object given an uttered word. This regret-minimization approach (Auer *et al.* 1995) is similar to Bayesian updating of the conditional probability of attribute values given words. As an online learning method, it offers performance guarantees for any given stopping time and noise. The vectors of weights form the wubble’s representation of a learned concept.

These criteria are important in a learning method because all word learning in the Blocks Room happens during online interaction with a child. For this purpose, Wubble World offers a series of goal-oriented tasks, such as the construction

of a set of stairs to retrieve a hanging golden apple (depicted in Figure 2).



Figure 2: Wubble Room

Paul Bloom (Bloom 2001) describes the meaning of a word as a certain mental representation or concept combined with a form. For every word a wubble knows, it maintains a corresponding concept, which is the wubble’s mental representation of the meaning of the word, and consists of two parts: a probabilistic representation of each feature as it contributes to the wubble’s understanding of the concept, and the set of positive examples of the concept. This can be seen in Figure 3. Using this notion of meaning, given the scene in Figure 2 and the input sentence “go to the cylinder,” we can begin processing the sentence.

The first step is to parse the sentence into a semantic representation, roughly equivalent to the logical form (LF) of the sentence. In this LF, the verb “go” specifies an action, and, since we have assumed a basic level of motor competence, the Wubble knows how to execute this action natively. In contrast, the meanings of nouns, adjectives, and prepositions are not available in advance, and so the word “cylinder”, known to be a noun from the LF, will need to be learned. This learning is driven by regret minimization.

When a new word is encountered, the weight vector for each attribute is initialized with a uniform distribution of weights, resulting in a uniform probability distribution, which is also the state of maximum entropy for the distribution. This reflects the state where the Wubble has no evidence to determine the relative contributions of color, shape, and size to the meaning of “cylinder”. Moving from this initial state to a meaningful (low-entropy) state traditionally requires a large amount of random trial-and-error on the part of the agent. However, we are able to leverage the child’s power

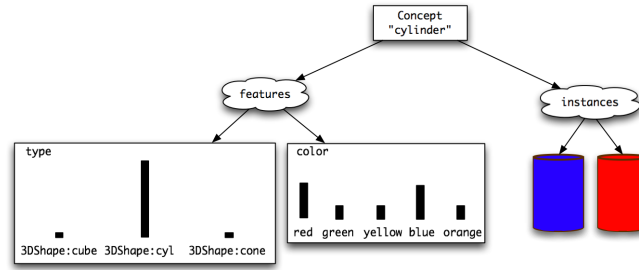


Figure 3: A graphical representation of the *concepts* stored by the Wubble.

of disambiguation, resulting in rapid convergence on the concept in question. Wubbles accomplish this by simply asking the child to point to the unknown object, in this case the cylinder. The attributes of the target object are then given more weight in the probability distribution. The wubble will continue to query the child on subsequent occurrences of “cylinder” until the entropy in the attribute distributions falls below a certain threshold. The result is that the Wubble will have a solid concept of cylinder after only hearing the word a few times.

After this point the wubble is able to act immediately on hearing the word “cylinder”, by selecting the object in the room that most closely matches its concept of cylinder. This is effectively a form of prototype reasoning, where the prototype is generated by stochastically sampling the probability distribution for each attribute, as defined by the weight vector and shown in Figure 4. The distance is then computed from the actual attribute values of each object in the room to this prototype set of attribute values. It is important to note that each attribute is weighted by the entropy of its distribution, so that high-information attributes are given more credibility than low-information ones. Thus, even though the prototype for “cylinder” might be *COLOR: blue, SIZE: small, TYPE: 3DShape:cyl*, the low entropy of the TYPE distribution and high entropy of the COLOR distribution would mean that a red cylinder would be favored over a blue cube, for example. This framework ensures that the wubble will act decisively, but because the underlying probabilities are updated so as to minimize regret, it will still perform limited exploration (the amount of exploration can

be increased via negative feedback from the child).

A further advantage of this representation for natural language understanding is that it allows the wubble to fluidly combine the meaning of a noun with the meaning of an adjective to understand an entire noun phrase. In fact, simple adjectives can be represented using the same formalism as nouns, and so the same learning techniques can be applied. The prototypes corresponding to each word are then combined to find the object in the room that most closely matches the meaning expressed in the entire phrase. In principle, this method can be applied to an arbitrary number of adjectives and/or nouns.

## Summary and Discussion

Situated language data isn’t hard to come by, but language data accompanied by a precise encoding of the world state at the time of the utterance is. In this paper we introduced the Wubble World project, which we built to address this problem by harnessing the creative energy of children. The social interaction and games in Wubble World will allow us to collect a language/scene corpus significantly larger - we hope - than those that exist today. This corpus will then be used to bootstrap a more sophisticated language learning effort than the system currently employs.

A wubble can learn some of the things a small child can learn, in much the way a small child would learn them: through interaction with an “expert” who already knows the language. Our regret-minimization approach to concept acquisition provides results that appear to mimic fast-mapping data from child language acquisition. With this technique a wubble learns to structure aspects of its

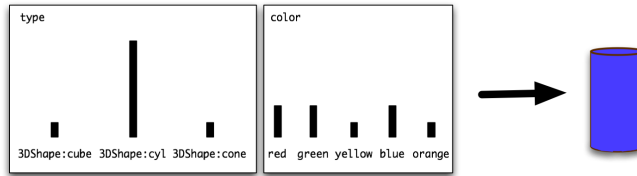


Figure 4: Creating a prototype from the representation of a concept

environment, in particular object traits like color, shape, and size.

If the wubble’s sensing seems simple, it’s because it is. Our goal was not to simulate the complexities of real visual processing; rather, we have taken our cue from biological systems and assumed that “real” sense data has been abstracted by lower-level systems into a higher level representation. It is these representations, and not continuous signals, that the wubble uses to learn concepts. This approach seems promising for domains that can be reasonably discretized; working directly with continuous data is a subject of further study.

In any social language environment there is a danger that utterances will either be nonsensical or require extra-scene context to understand. This is an unavoidable problem, and one that makes language understanding such a difficult goal. We address this principally by trying to skew interactions in Wubble World to make what’s happening in the scene - and thus, what can be automatically captured - salient enough to attract comment from the wubbles. This is more art than science, and the process thus far is crude. It is likely that a large portion of the language used will simply be impossible to pin onto any particular referent. We hope to address the issue of “knowing when we don’t know” in subsequent work.

Another shortcoming of our system is that the language model, while sufficient to learn words and concepts, is uncomfortably *deus ex machina*: by using a parser we have endowed our wubbles with a simple universal grammar. How would this primitive syntactic competence really be acquired? We think Wubble World provides a unique environment to explore this problem.

Much remains to be done. The platform is young, the system complex, and the challenges, both scientific and logistical, are substantial. We

have only just begun to tap into the domain of social interaction, and our first attempts at solutions have unearthed many more problems, not the least of which is a very practical issue: if children don’t want to play in Wubble World, we don’t get any language data. It’s no easy thing to design a fun game, let alone a game fun enough to offset the decidedly awkward aspects of natural language communication over the internet. We’ve thought a lot about how to structure interactions in Wubble World to maximize the fun factor and how to elicit the most meaningful language data. The system described in this paper is the beginning of that effort.

## References

- Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 322–331.
- Bloom, P. 2001. *How Children Learn the Meanings of Words*. MIT Press.
- Gorniak, P., and Roy, D. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research* 21:429–470.
- Lakoff, G., and Johnson, M. 1980. *Metaphors we live by*. University of Chicago Press Chicago.
- Roy, D., and Reiter, E. 2005. Connecting language to the world. *Artificial Intelligence* 167(1):1–12.
- Winograd, T. 1972. *Understanding Natural Language*. Academic Press, Inc. Orlando, FL, USA.