Grounded Pronoun Learning and Pronoun Reversal

Kevin Gold
Department of Computer Science
Yale University
51 Prospect St.
New Haven, CT, USA

Brian Scassellati
Department of Computer Science
Yale University
51 Prospect St.
New Haven, CT, USA

***Abstract*— An embodied language-learning system is presented that can learn the correct deictic meanings for the words "I" and "you." The system uses contextual clues from already understood words and sensory information from its environment to determine the most likely grounding for a new word. The system also serves as a model for the phenomenon of pronoun reversal among congenitally blind children, as the system learns that "you" is its own name when it is blinded. The system is novel among grounded systems in that it learns language by observing interactions between other agents, rather than from a helpful caregiver, and in that it associates words with social roles rather than reasoning about visual appearance alone.**

***Index Terms*— pronouns, functional language learning, deixis, pronoun reversal, humanoid robot, grounded language, blind language acquisition**

I. Introduction

The pronouns "I" and "you" present an interesting challenge to computational models of word learning. Though the emphasis in word learning research has been squarely on one-on-one interactions between parent and child, the word "you" cannot be learned through this kind of interaction alone, because "you" only ever refers to one entity in that situation: the child. How is the child ever to learn the more general meaning of "you" from this kind of interaction alone?

Oshima-Takane has argued [1]–[4] that "I" and "you" are not learned solely through one-on-one interaction, but by observing others interact with each other. The evidence includes the fact that second-born children learn "I" and "you" faster than firstborns [2] and that a neural network in simulation learned "I" and "you" faster when there were more participants to observe [4]. Though neither of these findings is conclusive evidence in itself, they both support the intuition that there is little to distinguish "I" and "you" from proper names if the child only receives input from the primary caregiver.

Learning by observing other agents interacting with each other presents its own set of challenges, however. One cannot depend on the participants to point at each other as they say "I" and "you," as this kind of gesture is unnecessary (if not rude) in normal conversation. Without the helpful caregiver tailoring the learning experience for the child, much of the proposed scaffolding for language learning, such as the exaggerated prosodic cues of "motherese" or helpful finger-pointing, no longer applies. The only model to date for learning "I" and "you" ignored this problem entirely by treating the referent of the word as a direct input to the network: an integer that gave the exact identity of who the sentence was about [4]. Thus, realistic models for pronoun learning through observation are fairly unexplored territory, and may provide insight into how children can learn other words through observation.

Realistic models of pronoun learning must also account for *pronoun reversal*, a phenomenon that occurs primarily among autistic children [5], linguistically precocious children before the age of 2 [6], and congenitally blind children before the age of 5 [7], [8]. Pronoun reversal is the usage of "you" or another pronoun where "I" is meant, or vice versa. For instance, a child exhibiting pronoun reversal may say "I don't want to comb your hair" when her mother offers to comb her hair [8]. Though much pronoun reversal can be attributed to echolalia or imitation, particularly in the case of autistic children, at least 48% of pronoun reversals in one study of linguistically precocious children were not imitations [6]. Another study that included a congenitally blind pronoun reverser found that only 29.9% of his reversals were attributable to imitation [9].

Pronoun reversal among blind children has been attributed to a poorly developed understanding of self [7], deficits in perspective-taking [8], and even partial autism brought on by an inability to see facial expressions [10], but none of these explanations has been particularly supported by the evidence. Instead, each study used the phenomenon itself to justify its interpretation of the blind children's behavior. If pronoun reversal is primarily a problem of language instead of concepts, as Oshima-Takane has argued [1], such "social deficiency" interpretations may be doing blind children a disservice.

In addition to its psychological interest, the problem of learning "I" and "you" is interesting from an artificial intelligence point of view because it challenges common assumptions about how robots should learn language. One common assumption is that robots learn language through interaction with a single person at a time, who is talking directly to the robot in a situation of shared attention [11], [12]. This is clearly insufficient to learn that "you" refers in general to the

person being addressed, because the teacher can only refer to the robot as "you," giving the robot no reason to believe that the word can refer to agents besides itself. Designing a robot able to learn from conversations not directed toward it is a daunting proposition, however, because the robot cannot rely on controlled gaze direction and pointing to determine reference. Previous robotic language-learning methods have also been tailored to learn words associated with properties that are computable from the image itself [11], [13], [14], but "I" and "you" are not associated with any particular visual properties.

Oshima-Takane's theory of pronoun learning suggested to us that robots should learn "I" and "you," and by implication other words as well, by watching humans interact with each other, rather than by interacting with a teacher. On the other hand, Oshima-Takane's pronoun-learning simulation [4] did not deal with the problem of reference in the real world, nor did it take into account the existence of irrelevant properties that might be accidentally associated with the pronouns instead. In short, this appeared to be a fruitful area for an interdisciplinary approach that could inform both sides.

We have implemented on a physical robot a method that successfully learns the semantics of the pronouns "I" and "you." Though we have presented this system once before [15], the current work expands on that work in the following ways. First, it was unclear how fast the learning was taking place, and what affected this rate; we analyze this here. Second, in our earlier study it was not clear what the competing word hypotheses meant, since these were essentially fake variables assigned randomly to the sensed agents. Though there are now fewer competing hypotheses, they now have clear interpretations. Third, we have now implemented our model for blind pronoun reversal on the robot. Finally, while we originally presented these results to an audience mostly interested in robot usability, we felt the results would be more interesting to an audience interested in learning and development.

II. A Method for Bootstrapped Word Learning

The challenge in learning words for pronouns is twofold. First, how can the listener deduce who the word is about – the *referent* of the statement? Second, how can the learner determine which *property* of that person the word refers to – whether it refers to that person's proper name, conversational role, current action, physical appearance, or some other property?

We do not require that the robot begin with no vocabulary whatsoever. To do so is an unnecessary handicap, both for modeling purposes and for a practical implementation. Children typically do not learn pronouns until they are roughly 2 years old, by which time they have already learned several concrete nouns and verbs [16]. Moreover, if we wish to build robots that can learn language from their environment, we should be more interested in the "inductive step" of adding to their vocabulary than the "base case" of learning first words, since groundings for some words can be programmed in before run-time. Thus, there is no reason to require that the robot begin with an empty vocabulary.

With even a small vocabulary, the problem of determining referent becomes easier, because the robot can use the words it hears as pointers to objects in the environment. This use of context is especially important because a statistical method that associates every word with everything in the robot's environment cannot learn that "I" refers to the speaker and "you" refers to the addressee. During each utterance, there is always a speaker and always an addressee. The words "I" and "you" must therefore occur equally often in the presence of both speakers and addressees. Only the use of the context-fixing phrase "got the ball" allows the robot to concentrate on which of these two agents the sentence is actually about.

But there still could be any number of other properties about that agent to which the unknown word might refer. For all the learner knows, the speaker could be talking about the color, position, or size of the person that has the ball. Thus, while determining the referent is *necessary* to learn the words "I" and "you," it is not *sufficient*, because there is still the question of what property of the person made the word applicable.

For this part of the learning, we fall back to statistical methods, to find which properties are most strongly associated with which words. Once reference has been established by other words in the sentence, the system uses 2×2 chi-square tests [17] to find statistically significant associations between the spoken words and properties of the referent. The four squares of the chi-square table correspond to the cases of whether a word is present or not in the sentence, and whether the property is true of the current referent. Chi-square tables and tests for every word-property pair are tabulated and computed; the resulting p-values can then be interpreted as confidences in the word-meaning pairs. (Chi-square values resulting from a *lower* than expected rate of coincidence are ignored.) Finally, because many words and properties that are only tangentially related may become significant with enough data, only the highest chi-square value is taken to be the word's meaning. This method is similar to that used for finding statistically significant word collocations in text [17], only here words are matched to sensed properties instead of other words.

Figure 1 provides an example of how the system analyzes an utterance.

III. Experiments

The algorithm described above was implemented on Nico, our humanoid robotic research platform (Figure 2), in the context of a game of catch. Vision was handled by one of the robot's cameras running at 320x240 resolution. This image

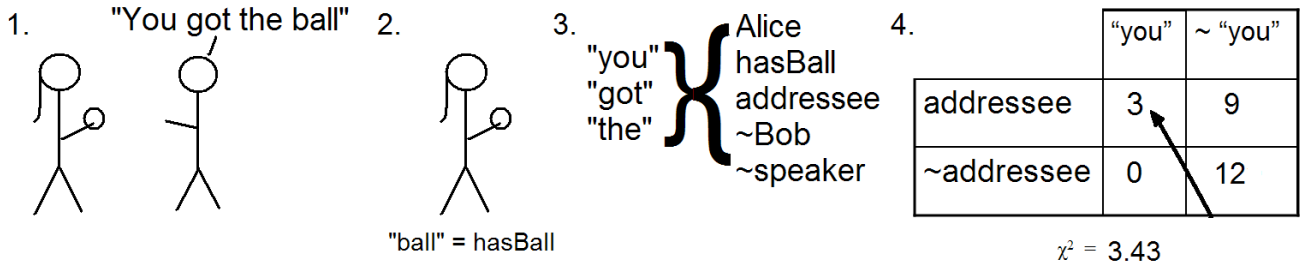

Fig. 1. An example of how the system associates words with properties. (1) Bob says to Alice, "You got the ball." The speech recognition software turns the speech into a string, while localization determines that Bob is the speaker and Alice is the addressee. (2) The system searches for words it already understands, and finds that "ball" corresponds to the hasBall property. The system designates Alice as the referent for the remaining words, because she has the ball. (3) Each word that was not understood is associated with Alice's properties, by increasing the words' collocation counts with those properties. (4) The updated collocation counts are placed in 2×2 chi-square tables to compute the significance of each word-property association.

Fig. 2. Nico, the robot on which the system was implemented.

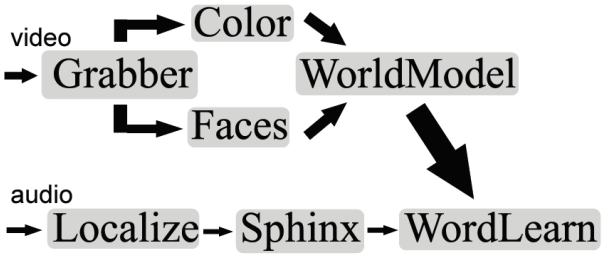

Fig. 3. Processing pipelines for the word-learning robot. A frame-grabber passed individual frames to a color processing module that found the bright yellow ball, and a face detection module that found the subjects' locations. Audio was localized between two microphones to determine speaker, then passed to the Sphinx-4 speech recognition program. The recognized speech was then compared to the sensed environment to ground word meanings.

was then passed to a module running the Intel OpenCV face detector and another module built to find the bright yellow ball in the image. Possession of the ball was determined by which face was closer to the ball's centroid.

In the audio pipeline, a dual-channel microphone set was used to determine who was speaking. The two microphones were set roughly 30 cm apart, and roughly 40 cm from each speaker. This separation was sufficient to localize the auditory signal to "left" or "right," depending on which audio channel exceeded a threshold first. Audio was then passed to the Sphinx-4 speaker-independent speech recognition system, which turned the words to text via context-free grammar. The result was then passed to the word learning module for association (Fig. 3).

Processing was performed in real time, with the primary bottleneck being the speech recognizer. The recognizer's self-reported processing time was 1.56 seconds per utterance. Ball detection occurred at roughly 30 FPS, while face location was updated at a rate of 1 FPS. Time to calculate chi-square values was negligible.

The world model consisted of the following salient properties: *speaker*, which was true of the person speaking; *addressee*, which was assumed to be the other person; *hasBall*, which was true of whoever had possession of the ball; and a unique property for each agent to represent identity. The robot's existing vocabulary consisted of knowing that "got" referred to the *hasBall* property, allowing the system to understand that either "got it" or "got the ball" referred to the ball. (The same results could have been achieved by restricting subjects to saying "I got the ball" and "you got the ball.") The experimenter and a fellow graduate student tossed the ball back and forth and commented on the action by saying "I got the ball," "you got the ball," "I got it," or "You got it," for a total of 50 utterances.

The second experiment used the same setup as the first, but this time the robot could not see who had the ball. Instead, it could only sense when the ball was very close to it, using the size of the ball in its visual field as a proxy for the sense of touch. Thus, the robot could only tell when it itself possessed the ball. Also, though the robot could sense who was speaking, it could not tell who was being addressed.

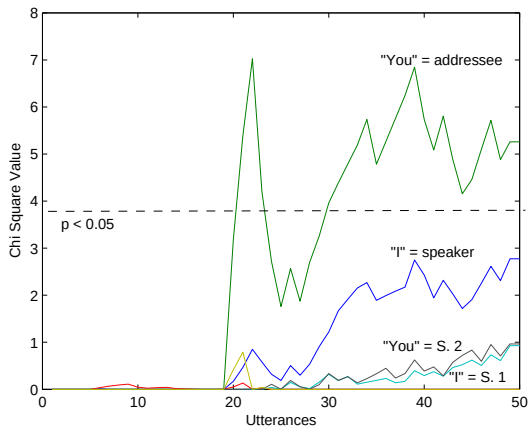

Fig. 4. Chi-square values for the word-property associations in Experiment 1. The large jump at utterance 20 marks the first time the system heard the word "you." The chi-square value for statistical significance (3.84) is given for reference, though meaning is attributed to a word based on its highest chi-square value. The second-best hypotheses, indicating that "I" and "you" are the names of the two subjects, are also shown.

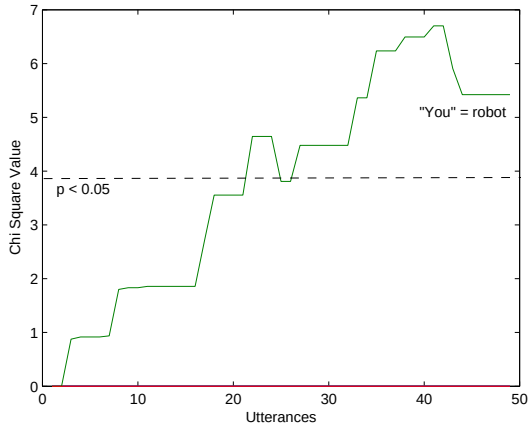

Fig. 5. Chi-square values for the "you = robot" hypothesis when the system was blinded. No other association achieved a valid chi-square value, though this value would hold equally for any other property that was uniquely and consistently true of the robot.

In addition to passing the ball back and forth, the two speakers could also pass the ball to the robot, and say either "You got the ball" or "Nico got the ball." Finally, because the statements "I got the ball" and "you got the ball" now provided no linguistic information to the robot when it did not have possession of the ball, names for the humans were added to the robot's vocabulary. The humans referred to each other by name when addressing the robot, and as "you" when addressing each other.

IV. Results and Analysis

As Figure 4 shows, the unblinded system showed clear long-term trends toward learning that "I" refers to the speaker, and "you" to the person being addressed. These chi-square values increased steadily over time, while the competing hypotheses that they were names for the individuals rose at a much slower rate.

For the first 19 utterances, the speakers had only used the phrases "I got the ball" and "I got it." This produced zeros in the denominator for at least one chi-square term in all of the relevant word-property associations. The system had no reason to assign any meaning to "I," because it was a part of every sentence, and no reason to assign any meaning to "you," because it was a part of none.

When "you got the ball" was finally spoken at utterance 20, the chi-square value for the association between "you" and the *addressee* property spiked. This was because both "you" and reference to the addressee were rare events so far, making their coincidence highly significant. On the other hand, the usage of "I" and reference to the speaker were both still very common events, and so little could still be concluded from their common occurrence.

This points to a rather surprising fact about chi-square word-object associations: the more common properties generate *less* confidence. Because saying "you got the ball" remained less common than saying "I got the ball" (and reference to the addressee less common than reference to the speaker), the confidence in the I/speaker association remained lower than the confidence in the you/addressee association over the course of the experiment.

Examination of the underlying chi-square equations confirms this analysis. Suppose word W always refers to agents with property X, and for the moment assume that there is no error in hearing the word or perceiving the property. Let w be the number of times W has been heard to refer to an agent with property X, and let p be the observed frequency ($0 < p < 1$) with which property X is true of a referent regardless of what words are heard. Let C be the total number of words that the system hears, and assume $w \ll C$ so that the contribution from irrelevant words in the absence of the property is small. Then it can be shown that

$$\chi^2 \approx w(1/p - p) \tag{1}$$

The calculations are somewhat tedious, and we omit them here; they require the approximation $(Cp-w)^2/p(C-w) \approx Cp - w$ in the chi-square term corresponding to the case of ($\sim$word $\wedge$ property). The chi-square value thus increases linearly with the number of times the target word is heard, but inversely with the frequency with which the property is observed to be true of a referent.

The drop in confidence beginning at utterance 23 was caused by a series of speech recognition and localization

errors. Since the chi-square equations quickly become unwieldy when multiple variables representing different error rates are added, we shall analyze only the most damaging kind of error here: the effect of mistakenly believing that the word has been used in conjunction with an agent that does not possess the correct property. This occurred in our experiment when the speech recognition software mistook one pronoun for another, or when the localization routine mistakenly attributed a correctly heard statement to the wrong person. If ϵ is the rate at which occurrences of the event (word $\wedge$ property) are mistakenly interpreted as the event (word $\wedge$ $\sim$property), then the analogous assumptions to the errorless case result in the expression:

$$\chi^2 \approx w(\frac{\epsilon^2}{1-p} + \frac{(1-\epsilon)^2}{p} - p - \epsilon) \tag{2}$$

Here, w is the true count of the number of times the word was used in conjunction with an agent bearing the correct property; the agent's count is actually $(1-\epsilon)w$. The reader can verify that when $\epsilon = 0$, equation 2 reduces to the errorless case of equation 1.

The dominant effect of increasing the error rate ϵ is in the term $(1-\epsilon)^2/p$, where ϵ has an effect inversely proportional to p. This partly explains why the decline in chi-square is so great due to error during the first few utterances of "you"; not only is ϵ effectively greater because of the small value of w, but the addressee property is uncommon, amplifying the effect of error. Asymptotically, however, the error merely changes the learning rate by a constant factor, leaving the rankings of word-property associations unchanged.

It has been suggested in the word-word collocation literature that chi-square results should be considered untrustworthy unless the expected values in each square of the chi-square table are at least 5 [17]. In this case, this heuristic results in the rule that judgment should be withheld until the following condition occurs:

$$w > \max(5/p, 5/(1-p)) \tag{3}$$

In our experiment, this would have resulted in the system withholding judgment on the word "you" until roughly utterance 32, thus avoiding the awkwardness of revoking and then reinstating confidence in the association with the addressee property. More balanced occurrence of the "I" and "you" cases, so that $p = 0.5$, would have resulted in confidence much earlier, around utterance 10.

In the second experiment, association of "you" with the robot's identity reached significance in about twenty utterances when the robot was blinded (Fig. 5). This chi-square value would hold equally well for any property that was always true of the robot when "you" was spoken, but never detected about other agents. Thus, if the system had been able to tell that it was the addressee when it was being addressed, the chi-square value for associating "you" with *addressee* would have been the same as that for the association of "you" with the robot's identity. To distinguish between the two hypotheses, the robot would then have needed to employ some other criterion besides strength of association. "I" was not associated with any property in the blinded case because the system had no way of determining the referent of the sentence "I got the ball" when it did not possess the ball.

V. Discussion

The approach presented here is more abstract than previous approaches to robotic word learning, which have usually focused on associating words directly with sensory properties [13]. In [11], for example, word associations were based on mutual information between phoneme sequences and slightly preprocessed representations of shape. While appearance is an important cue for many nouns, it is not clear that this approach can applied with success to the majority of words. For example, words for artifacts, such as "paperweight," are more often defined by their function than by their shape [19].

The words "I" and "you" are not associated with visual properties, but with agent roles. The fact that these words are universally learned with relative ease suggests that roles and actions may be at least as important to word learning as visual appearance. By making functional properties salient for word acquisition, words that would be very difficult to associate with raw sensory data may become much easier to acquire. In addition, functional definitions are more closely related to planning, which should prove useful in most applications of language.

The system presented here also demonstrates the utility of being able to use sentence context to narrow the space of possible referents for an unknown word. The conversational roles of "speaker" and "addressee" are always present in an observed conversation, but the system presented here only associates the role associated with the sentence referent with the unknown words. Without sentence context to determine referent, both properties would be equally associated with "I" and "you." The use of statistical evidence to determine the exact property associated with a word after reference has already been established can be seen as a middle ground between associationist models and "Augustinian" approaches [19] that rely on context and inference to determine reference.

The mathematical properties of chi-square values appear to be well-suited to word learning. Confidence in a word-property association is linearly proportional to the number of times the word has referred to the property, which is a reasonable learning rate. Moreover, if a property is common, the system is reluctant to assign a word to it, while uncommon properties are quick to be assigned words. This is useful behavior, because it allows the system to quickly associate words with the aspects of agents that are unique to them. The ability to quickly learn new words for new properties is called "fast mapping" in the developmental

literature [20]. Fast mapping is often explained as resulting from inference from contrast, as in the statement "Bring me the beige one, not the blue one" [21], but chi-square tests can be seen as containing contrast information implicitly using the learner's experience. The system's reluctance to assign words to common properties would make "I" and "you" slow to be learned compared to other words if it were learning a larger vocabulary, but this matches the developmental finding that "I" and "you" appear later than many nouns and verbs, despite their prevalence. Though children obviously do not consciously perform chi-square calculations, the underlying neural machinery is probably subject to the same constraints as our system if it is to find meaningful, rather than accidental, coincidences of word and property.

Any model of human pronoun learning should also be able to explain pronoun reversal. The present model displays some of the behavior of pronoun learning, but it does not yet explain all of it. There is anecdotal evidence that blind children have trouble understanding "I" [7], which is less easily accounted for by our model since they should be able to sense the deictic shift indicated by a change in speaker. Our model also does not account for the finding that pronoun reversal is more common in utterances that involve more than one pronoun [6]. Some cases of pronoun reversal have involved the substitution of third-person pronouns, instead of "you," for "I" [8], a case that we do not handle here. Finally, autistic pronoun reversal may indeed stem from theory of mind deficits instead of linguistic error. On this last question, our previous study [15] provided suggestive evidence that theory-of-mind heuristics may be important for inferring the referents of many common statements.

The next steps for our research include the learning of other deictic pronouns, such as "this" and "that," and the extension of this framework to other word categories. The idea that being the subject of an agent's attention is a property that can be treated the same as a visual property was critical to learning the word "you," and we expect that a similar principle should hold true for other deictic pronouns when applied to objects. Nothing about the present method restricts it to learning deictic pronouns; in fact, the idea of attention adding a property to its target is specifically meant to make deictic pronoun learning compatible with learning words for more prosaic properties such as shape and color. It is our hope that by focusing on the hard case of learning deictic pronouns, the general principles of how humans learn the meanings of words will become clearer.

Acknowledgments

Support for this work was provided by a National Science Foundation CAREER award (#0238334). Some parts of the architecture used in this work was constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots) and from the DARPA CALO/SRI project. This research was supported in part by a grant of computer software from QNX Software Systems Ltd.

References

[1] Y. Oshima-Takane, "Analysis of pronominal errors: a case study," *Journal of Child Language*, vol. 19, 1992.

[2] Y. Oshima-Takane, E. Goodz, and J. L. Derevensky, "Birth order effects on early language development: do secondborn children learn from overheard speech?" *Child Development*, vol. 67, pp. 621–634, 1996.

[3] Y. Oshima-Takane, "Children learn from speech not addressed to them: the case of personal pronouns," *Journal of Child Language*, vol. 15, pp. 95–108, 1988.

[4] Y. Oshima-Takane, Y. Takane, and T. Shultz, "The learning of first and second person pronouns in English: network models and analysis," *Journal of Child Language*, vol. 26, pp. 545–575, 1999.

[5] C. Lord and R. Paul, "Language and communication in autism," in *Handbook of Autism and Pervasive Development Disorders*, 2nd ed., D. J. Cohen and F. R. Volkmar, Eds. New York: Wiley, 1997, pp. 195–225.

[6] P. S. Dale and C. Crain-Thoreson, "Pronoun reversals: who, when, and why," *Journal of Child Language*, vol. 20, pp. 573–579, 1993.

[7] S. Fraiberg and E. Adelson, "Self-representation in language and play," in *Insights from the blind*, S. Fraiberg, Ed. New York: Basic Books, 1977.

[8] E. S. Andersen, A. Dunlea, and L. S. Kekelis, "Blind children's language: resolving some differences," *Journal of Child Language*, vol. 11, pp. 645–664, 1984.

[9] M. Pérez-Pereira, "Deixis, personal reference, and the use of pronouns by blind children," *Journal of Child Language*, vol. 26, pp. 655–680, 1999.

[10] R. Brown, R. P. Hobson, A. Lee, and J. Stevenson, "Are there 'autistic-like' features in congenitally blind children?" *Journal of Child Psychology and Psychiatry*, vol. 38, pp. 693–703, 1997.

[11] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, pp. 113–146, 2002.

[12] L. Steels and F. Kaplan, "Aibo's first words: The social learning of language and meaning," *Evolution of Communication*, vol. 4, no. 1, 2002.

[13] D. Roy, "Grounding words in perception and action: computational insights," *Trends in Cognitive Sciences*, vol. 9, no. 8, pp. 389–396, 2005.

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

[15] K. Gold and B. Scassellati, "Using context and sensory data to learn first and second person pronouns," in *Human-Robot Interaction 2006*, Salt Lake City, Utah, 2006.

[16] D. J. Messer, *The Development of Communication*. West Sussex, England: Wiley, 1994.

[17] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[18] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[19] P. Bloom, *How Children Learn the Meanings of Words*. Cambridge, Massachusetts: MIT Press, 2000.

[20] S. Carey, "The child as word learner," in *Linguistic Theory and Psychological Reality*, J. Bresnan, G. Miller, and M. Halle, Eds. Cambridge, MA: MIT Press, 1978, pp. 264–293.

[21] T. Heibeck and E. Markman, "Word learning in children: An examination of fast mapping," *Child Development*, vol. 58, pp. 1021–1034, 1987.