Using Context and Sensory Data to Learn First and Second Person Pronouns

Kevin Gold
Department of Computer Science
Yale University
51 Prospect St.
New Haven, CT, USA
kevin.gold@yale.edu

Brian Scassellati
Department of Computer Science
Yale University
51 Prospect St.
New Haven, CT, USA
scaz@cs.yale.edu

ABSTRACT

We present a method of grounded word learning that is powerful enough to learn the meanings of first and second person pronouns. The model uses the understood words in an utterance to focus on the agents to which they refer. The method then uses chi-square tests to find significant associations between the remaining words and the attributes of the relevant agents. We show that this model can learn from a transcript of a parent-child interaction taken from the CHILDES database [22] that "I" refers to the person who is speaking. With the additional information that questions about wants refer to the addressee, the system can also learn the meaning of "you" from observed dialogue. We show that an incorrect assumption about the probable referent of "want" questions can lead to pronoun reversal, a linguistic error most commonly found in autistic and congenitally blind children. Finally, we present results from a physical implementation on a robot that runs in real time. Our preliminary results on the robot, while indicative of the difficulty of using real sensors for concept learning, show that the model does work in practice.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: [speech recognition and synthesis]; J.4 [**Social and Behavioral Sciences**]: [psychology]; C.3 [**Special-Purpose and Application-Based Systems**]: [real-time and embedded systems]; I.2 [**Artificial Intelligence**]: Robotics

General Terms

Experimentation

Keywords

Natural language, grounding problem, pronouns, collocations, real-time, robot, word learning

1. INTRODUCTION

The learning of first and second-pronouns presents a puzzle for models of word acquisition. Suppose an experimenter is attempting to teach a robot some words. "Ball," the experimenter says, showing the robot a ball. After some repetition of this word, either alone or embedded in other phrases, the robot might associate the word with the object, and be able to point to the ball and say, "ball." A similar associative process was carried out by Roy and Pentland [1], who showed their robot Ripley a series of objects in conjunction with a recording of a mother teaching her child the words for the object. By finding the object-word pairings with the highest mutual information, Ripley learned words that corresponded to each object.

But now consider the case of "I" and "you." Suppose the experimenter points to himself and says, "I." Under the current state of affairs, the robot would point to the *experimenter* and say, "I." Conversely, if the experimenter pointed to the robot and said, "you," the robot would point to itself and say, "you." By assuming that these words are the names of things in the environment, the robot has accidentally reversed their meanings.

Pronoun reversal is not limited to hypothetical robots; humans can make the same mistake early in life. Children with precocious language development before the age of two often use "you" when they mean "I" [2]. Though it is not clear whether all children pass through this phase [3], two groups of children in particular are known for their tendency to switch first and second-person pronouns. The first group is autistic children [4]. The second is the congenitally blind.

It is perhaps unsurprising that autistic children should have trouble with "I" and "you," since they have such pervasive problems with other aspects of interpersonal communication. Blindness, on the other hand, is a much more understandable deficit, and thus it is surprising that such a seemingly simple condition should manifest itself with such an obscure symptom.

Many explanations have been put forward for why congenitally blind children should have a higher incidence of pronoun reversal than sighted children. One of the first claimed that blind children have a less developed sense of self, on account of their inability to see themselves – but the author apparently found the explanation so self-evident that

she provided no quantitative evidence to support it [5]. A later study suggested that blind children lacked proficiency in "perspective-taking" [6], though it too lacked any experimental evidence to support the idea. Yet another attempted to link the two pronoun-reversing groups together, by claiming that the blind child's inability to see faces and the autistic child's inability to understand them led to an impoverished social understanding [7]. These hypotheses all remained speculative, with little evidence to support any of them – though, to be fair, congenitally blind children are somewhat difficult to find before they enter elementary school, and small sample sizes have precluded statistically significant results [8]. Still, what quantitative evidence existed in these studies typically did not support one model over another, and none of them even acknowledged the fact that excessive pronoun reversal does not occur in all congenitally blind children, and possibly not even in a majority [8].

A more sensible explanation may be that the blind children in question simply did not receive enough data to give a correct interpretation. Oshima-Takane has argued that it is very difficult to correct a child's pronoun reversals, since the correction itself is subject to misinterpretation [9]. Rather, the child must learn what "I" and "you" mean by observing others. Oshima-Takane has shown that children with older siblings learn correct pronoun usage faster than eldest children [10]. She also found that children will correctly use first-person pronouns earlier if parents demonstrate the correct usage by saying "I" to each other while pointing at themselves [11]. (A similar experiment in which parents demonstrated "you" produced inconclusive results [11].) Furthermore, a neural network model began to correctly produce "I" and "you" when it was exposed to multiple demonstrators speaking to each other, but not when it was only exposed to one person's usage [12]. These results suggest that the blind children may simply be slow to learn correct usage because they lack the visual input that would clarify whom the speaker is addressing in multiple-person scenarios, and thus they simply lack a body of clear evidence from which to surmise the meaning of the word.

Oshima-Takane's neural network model was instructive, but it was not intended to be a practical system. One problem with Oshima-Takane's neural network model was that the inputs were heavily abstracted. Each person in the simulation was represented by a unique number, and the training inputs were the three numbers indicating the speaker, the addressee, and the referent of the word. The system merely had to learn to produce "I" when the speaker number was equal to the referent number, and "you" when the addressee equaled the referent. But how would a real system identify these characteristics?

The identification of who is speaking seems straightforward, barring the usual problems of noise: watch for moving mouths, or localize the direction of the sound. The addressee could be determined by using cues such as head orientation and gaze direction [13].

It is determining the referent of a word that is most problematic for a real system. The speaker's gaze direction is of little help, because the speaker will look at the addressee regardless of whether he refers to the addressee or himself. Similarly, pointing at the referent is simply not realistic for first and second-person pronouns; pointing the addressee is not polite, and pointing at the self is usually unnecessary. If we are to believe that learning the correct usage of "I" and "you" requires observing other people, then we cannot rely on unnatural gestures.

Even if the referent were easily identified, any model that learns *only* the words "I" and "you" is effectively useless. In the real world, word learning is not a multiple-choice test. These words could be bound to anything, not merely functions of the speaker, addressee, and referent; how is the infant to know that "I" does not refer to a specific person, or to the speaker's gender, or to the color of the addressee's hair? A model of learning "I" and "you" must take into account the possibility of other meanings. Conversely, by thinking about how to solve this difficult problem, we may learn something about how children can learn new meanings for words in general.

We propose that the words "I" and "you" are actually learned in the context of complete sentences, and that the interpretation of other words in the sentence provides potential referents for the unknown words. By finding objects in the environment that match the known words in the sentence, the unknown words can then become associated with those objects and their properties.

2. WORD LEARNING FROM CONTEXT

Finding word boundaries within the auditory channel is beyond the scope of the current paper. In the present study, we shall treat the auditory channel as if it were plaintext. (In our implementation, described below, we made use of the Sphinx 4 speech recognition system [14], while the simulations were run on transcripts.)

We here model simple properties of people as boolean attributes. These attributes represent actions such as "speaking," or descriptions such as "blue." We do not elaborate in this model how the boolean representations of such attributes might form, though self-organizing maps seem reasonable [15]. Though boolean attributes may seem to be an oversimplification, it is worth remembering that Boole originally conceived of his logic as describing human rules of thought [16], and that boolean models of concept learning are still not entirely out of fashion [17].

Our basic method consists of two parts: an attention-focusing mechanism and a statistical learning mechanism. When a sentence is first heard, the words with understood meanings are used to determine what items in the environment are being talked about. All items in the environment that match the description are placed in a candidate pool of targets for the remaining words of the sentence. The positive attributes of these items are then associated with all of the remaining words in the sentence.

While Hebbian learning [15] may be a more biologically motivated mechanism of association, we chose a statistical implementation to allow a more principled threshold for association significance. Each word's probability $P(word)$ is estimated using the frequency of the word in the corpus, and each attribute's probability $P(att)$ is estimated

by the number of spoken words during which the attribute is true for some attended object. The null hypothesis is that these probabilities are independent, so that $P(word \wedge att) = P(word)P(att)$. Using the real frequency of the event $(word \wedge att)$, we can then use a chi-square test to determine whether the word is spoken in conjunction with the attribute more frequently than chance would dictate. This approach is essentially that used for word collocations in statistical natural language processing [18], but we apply it here to word-attribute pairs instead of word-word bigrams.

Just as in corpus-based statistical natural language processing, we expect that many attribute collocations will be significant for a single word if we are given enough data – not just the attribute that best captures the meaning. For this reason, we rank the chi-square values for each word, and accept only the association with the highest chi-square value. In addition, chi-square values that do not achieve significance (here, $p < 0.05$) and those representing negative correlations of word and attribute are discarded. Finally, we ignore associations for which we have insufficient data; following a convention for chi-square analysis, this means ignoring associations with chi-square tables that have an expectation in any of the cells of less than 5 [18]. (One advantage of using chi-square tests over mutual information, as was used in [1], is that mutual information exaggerates the importance of collocations with sparse data [18].)

When we are done, we are left with a few words whose meaning can be reliably interpreted from observation. We assume that the part of speech can be inferred from another mechanism that handles grammar, which we do not model here – so a word associated with the attribute "speaker" may be a verb such as *speak*, an adjective such as *talkative*, or even a pronoun: *I*. We hypothesized that by associating the word "I" with the attribute of *speaking*, a child or a robot can learn that "I" refers to *the person that is speaking*.

Similarly, we hoped that the word "you" could become associated with the action of attending to a speaker. We suspected that our hypothesis might run into a pragmatic problem, however: it is rare to tell people obvious facts about themselves (e.g., "You are wearing a blue shirt"). We discuss our findings, and our solution to this problem, below.

3. EXPERIMENTS IN SIMULATION

3.1 Learning "I"

To test our model before implementing it on a real embodied system, we ran it on a transcript of a mother and her child playing catch [19] [20] [21], taken from the CHILDES database [22]. Only the raw words were used from these transcripts, and not the CHILDES part-of-speech annotations. We did not omit any "stop words" from the analysis. The corpus was relatively small, consisting of 1707 words. Of these, only 372 appeared in sentences with understood referents.

Because the transcript contains no stage directions, we can only infer the participants' actions from the text; luckily, the mother's comments make this process of interpretation fairly uncontroversial (e.g.: "You got it!", "Why are you blowing on it?", etc.). Annotating the text in this way produced six attributes that changed over the course of the text:

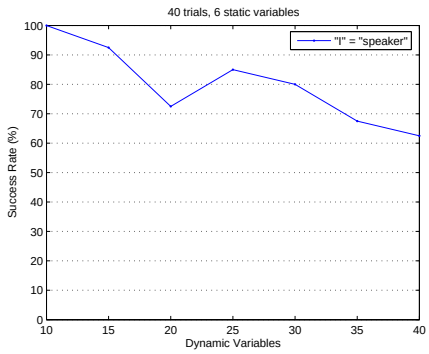

Figure 1: Success rates in learning "I" for varying numbers of speaker attributes that changed over the course of the simulation. No trial correctly surmised the meaning of "you" under these conditions.

throwing, catching, missing a catch, getting hit on the head, blowing on an object, and falling down. Accordingly, we assumed that the system already knew the words for these actions, as well as the word "mommy" and the name of the child.

To avoid bias in the annotation process, and more fully model a complicated environment, we added additional dynamic variables which did not correspond to anything in the script, but that changed with probability 1/2 from line to line. These represented other attributes of the attended objects that were changing, and could potentially be associated with the dialogue by accident. (The number of these distractors varied according to experimental condition, as described below.) In addition, we added six random attributes to each actor that remained fixed over the course of the interaction. We did not vary the number of static attributes, since all of them that were true for an individual would have an equal chi value for each attribute. Binding a pronoun to a static attribute that was true for one actor but not the other would indicate that the system had failed to generalize its meaning across individuals.

Finally, we added two attributes *speaking* and *addressee* that were true or false depending on who was delivering the line of dialogue.

A trial was considered a success if, by the end of the exchange, the word "I" was associated with the attribute *speaking*. Figure 1 shows the success rates out of 40 trials for varying numbers of dynamic variables. Even with large numbers of distractor attributes, the system performs well given the small size of the corpus. (We shall see later that increasing the number of understood sentences produces better results.)

Remarkably, this method succeeds even though one of the participants is occasionally lapsing into pronoun reversal:

CHILD: I got it!
MOTHER: Mommy got it.
MOTHER: That's right, mommy got it.
MOTHER: You didn't get it.

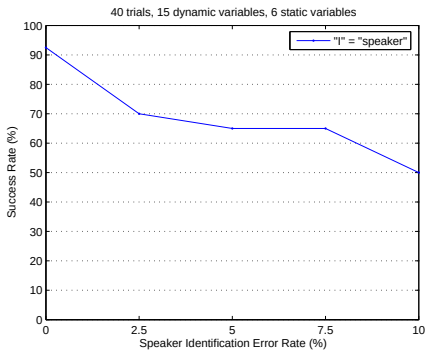

Figure 2: Success rates for learning "I" with simulated speaker identification errors.

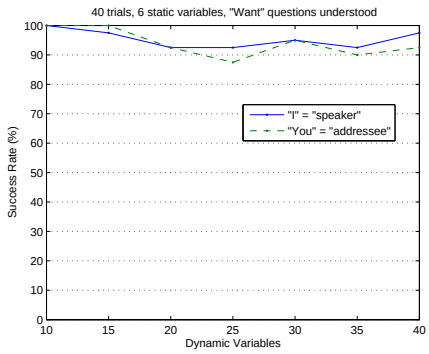

Figure 3: Success rates in learning "I" and "you" when the system was allowed to infer that questions about wants referred to the person being asked.

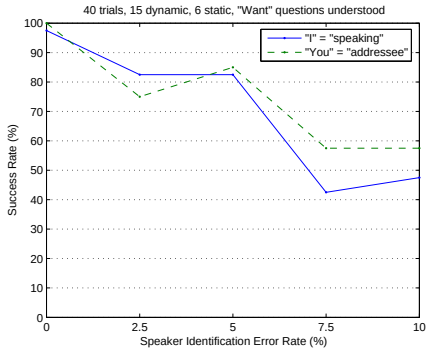

Figure 4: As figure 3, but with varying rates of speaker identification error.

MOTHER: Mommy got it. [19]

It is also worth noting that these trials produced very few erroneous associations. In the 6 dynamic variables case, only two words were given new associations: "I" was associated with speaking, and "it" was associated with catching (presumably because of the frequency of the phrase "I got it" in the script). Even in the 40 dynamic variable case, only 10 (25%) of the trials produced more associations than this, associating "you" or "the" with arbitrary dynamic variables.

But what if there is some confusion about who did the speaking? Though this may not be a problem for the human perceptual system, tolerance of such error may be a real concern for the roboticist. To simulate error in speaker identification, we added a chance for each statement that it would be attributed to the wrong person in the exchange. As figure 2 shows, even for small error rates, the chance of correctly learning the meaning of "I" falls off dramatically. Thus, even if our system is a good model of how humans learn the meaning of "I", we may expect some technical hurdles on the way to implementing it in an artificial intelligence.

3.2 Learning "You"

In none of the above trials did the system correctly learn the meaning of the word "you." We hypothesized that this was because statements about the observable properties of others are much less frequent than questions about others' non-observable properties – namely, their desires and mental states. (In our script, the exclamation "you got it!" does appear, but much less frequently than "I got it.")

Therefore, we added to the system the knowledge that questions involving the word "want" are about the person being asked. In terms of our model, this meant causing the word "want" to add all addressees to the list of potential referents of the sentence; their properties would then become associated with the remaining words in the sentence.

Our results are shown in Figures 3 and 4. The correct binding for "you" was correctly learned even for large numbers of distractor variables, and its success rate was comparable to that of "I" when speaker confusion occurred.

In addition, extraneous dynamic variables were not as detrimental to the learning of "I", because the additional sentences that could now be comprehended provided additional statistical evidence for the "I" hypothesis, and evidence against association with other dynamic variables. (The number of words in sentences with at least one known word increased from 372 to 417.) This suggests that as vocabulary size increases, conversations can be used more efficiently to test the meanings of words.

3.3 Pronoun Reversal

In the previous section, we added to the system the knowledge that questions about wants generally refer to the addressee. Though we do not model how this knowledge would be learned, we suspect that this process may be particularly error-prone in the case of autistic or blind populations. While autistic individuals can understand the concept of "want" regarding other individuals, they often have difficulty understanding situations in which one party has knowledge that the other doesn't [23]. Congenitally blind children do not perceive the head orientation and gaze direction cues that would allow them to see who is being addressed, and thus may make mistakes in understanding who is being asked. In either case, we might expect that these populations may sometimes incorrectly learn that "want" questions always refer to their own selves, since these are the learning instances with the clearest feedback. What happens to the meaning of "you" when this new rule is used for the

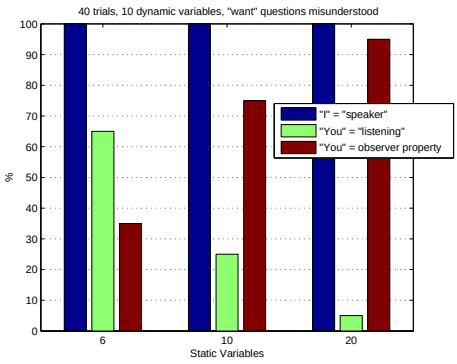

Figure 5: Response of the system when it assumes that questions about wants are always about itself. Given the chance to associate "you" with a property unique to itself, it does so; but if all attributes are shared with at least one other party, as is the case with small numbers of static variables, the word "you" gets bound instead to the property of "listening."

Figure 6: The robot on which the system was implemented.

interpretation of "want"?

We added a third observer child that represented the learner to our simulation. No changes were made to the script, but the learner now assumed that the word "want" referred to neither of the conversation's participants, but to itself. In addition, since there was now a distinction to be made between the addressee and other non-speakers, we added the attribute "listening" that was true for both non-speakers, while "addressee" was true only of the person being addressed. (Running this version on the previous experiment still produced "addressee" as the binding of "you.")

As Figure 5 shows, the correct learning of "you" dropped dramatically. When the number of static variables was increased to sufficiently high levels to ensure that at least one was uniquely true of the observer, the observer assumed that the word "you" referred to this property. In other words, when the system assumed that "want" questions were always directed at itself, it learned that "you" always referred to itself as well.

When there were no such unique properties, however, the next best hypothesis was that the word "you" referred to the "listening" property, which was true of anyone that was not speaking. Despite some evidence to the contrary in the script (e.g., "you got it," which would still associate you with only the catcher), none of the trials correctly associated "you" with the addressee alone.

Increasing the number of static attributes had no effect on the previous iterations of the experiment.

4. PHYSICAL IMPLEMENTATION

4.1 Methods

In this experiment, we tested whether a physical robot (Figure 6) could learn that "I" referred to the speaker by observing a game of catch, similar to the one described in the simulations above. The robot is the same size and shape as a one-year-old child, and is our platform for several avenues of developmental robotics research.

For this experiment, we used one of our robot's two high-acuity foveal cameras and a two-channel microphone as sensors. The camera ran at 320x240, and the resulting video was processed by a pipeline of four nodes running the real-time operating system QNX. These processed the image using the Intel OpenCV computer vision library [24] to find faces. A module for detecting highly saturated color pixels was co-opted for finding the ball; possession of the ball was determined by which face was closest. Though this method for sensing possession of the ball was decidedly tailored toward the experiment, it serves the purpose of illustrating the word learning model well enough.

The Sphinx 4 [14] speech recognition module ran on a Windows XP computer across the room, connected to the robot via a wireless connection. The Sphinx code was modified so that in addition to returning the interpreted speech, the speaker was also identified as being to the left or right of the microphones, based on which microphone the speech sound wave reached first. By matching the audio channels to the visual image, the speaker of each utterance was identified with very high accuracy – though this was partly because the speakers needed to be within a few feet of the microphones for the third-party speech recognition module to function. Figure 7 summarizes the modules involved in the physical implementation.

The words "got" and "caught" were given explicitly to the system to refer to the property of possessing the ball. For the Sphinx language model, we used a small (15-word) context-free grammar that included statements of possession ("you got it"), statements of desires ("I want the ball"), and some miscellaneous exclamations ("you blew it"). The full context-

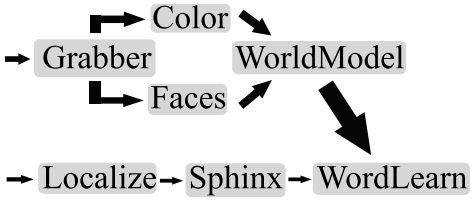

Figure 7: The modules used in implementing the word-learning method. An image grabber passed video to a color-saliency filter and the OpenCV [24] face-detection module, which then passed their results to the world model. Audio was first localized based on the disparity between the two channels, then passed to Sphinx [14] for speech recognition, and finally to the word-learning module for association with the properties of objects.

free grammar was as follows:

```
<S> := <catch> | <util>
<catch> :=  <subj> <verb> <obj>
<subj> := (Do you) | I | You | Kevin | Nico | Fred
<verb> :=  got | caught | blew | want
<util> := done | yes | no
<obj> := (the ball) | it
```

All processing occurred online and in real time, with a frame rate of 30 FPS for the ball detector, roughly 1 FPS for the face detector, an average delay of less than a second per utterance for the Sphinx module, and a negligible amount of time to calculate chi values.

To make up for the simplicity of the system, additional dummy dynamic and static variables were added to the sensed environment to bring the total number of each to 10.

Subjects were instructed to play catch in front of the robot and comment on the action using relevant statements from the grammar. They were given no guidelines as to how often each phrase should be spoken, nor on how long to wait between utterances or throws. As a result, the experiment encountered many of the difficulties involved in real conversation environments, including overlapping speech and an environment that changed as speech recognition was taking place.

4.2 Results

After 50 interpreted utterances, the highest-ranked interpretation of "I" was indeed "speaking," but the results were far from significant ($\chi^2 = 0.25$). It is difficult to say whether this means that statistical significance is too high a hurdle for practical, human-like word-learning, or if instead demonstrates the importance of these methods given the unreliability of the data. (Though the uneven timing of the speech samples made it difficult to match Sphinx's interpretations with ground truth, 14% of Sphinx's interpretations were clear errors in that they were never spoken by either party.) In particular, the Sphinx transcript contained only 8 positive examples of "I" referring to the speaker. Running the experiment for 4 more trials showed that it was still improving: the χ^2 value for the association of "I" and "speaking" increased to 0.52, while the association of "you" with "listening" increased over these trials from 0.78 to 1.78 ($p < 0.18$). Thus, with time we would expect these figures to achieve significance, and we are still in the process of collecting data to confirm this intuition.

We have not yet integrated the interpretation of prosody with Sphinx, so the robot cannot yet directly recognize when a statement is a question. However, we can structure the grammar so that beginning with "do" signals a question (e.g., "do you want the ball?"). While this is not a practical means of detecting questions in general, we can use this workaround to ask what happens when questions about wants are assumed to be directed toward the robot. As expected, performance does indeed worsen considerably: after 60 statements, neither "I" nor "you" were associated correctly, but were instead non-significantly associated with dummy dynamic variables. Because the "want" questions did not outnumber the statements of observable fact as they did in our simulation transcripts, the data supported neither the hypothesis that "you" referred to the robot, nor the hypothesis that it referred to the addressee.

5. CONCLUSIONS

5.1 Contributions

We see two major contributions that arise from this work. First, this is a potentially powerful system for learning grounded meanings for words in general – not just pronouns. Though the current implementation requires some words to already exist in the vocabulary in order to learn new ones, it should be straightforward to add initial noun learning to the system by means of an "object bias" that assumes the names for nouns are taught first, and thus most statistically likely upon presentation. (The "object bias" is a well-known empirical finding in children's language development; see [25] for a review.) Our method's use of context makes it more powerful than strictly associationist methods, which have been critiqued recently by Bloom [25]. We believe that approaches that combine statistical methods with pragmatic rules for interpreting the *intent* of utterances will prove most useful in robotic language learning.

Second, though it has historically been popular to attribute pronoun reversal to fundamental flaws in representations of the self [5] [7] [6], we hypothesize that this phenomenon is a problem of language learning, and not the child's underlying model of the world. Our results indicate that if a child assumes questions about wants are directed toward himself, that child will probably learn that "you" refers specifically to himself as well. This is not to suggest that autistic or congenitally blind children are naturally selfish. In the absence of cues such as gaze direction, it may be a natural assumption for a blind child that she is the default addressee. Autistic children, on the other hand, may have no difficulty determining who is being addressed, but may not make the pragmatic leap that questions about mental states should be directed toward the person in question, and not someone else. In either case, we may expect a delayed understanding that "you" can refer to others.

Our research further predicts that these populations could learn to correctly use pronouns more quickly if those around them were to use pronouns in sentences with clear referents. In particular, though mentalistic language appears to be the most common context of the word "you," using the word "you" to describe others in physical terms may help an autistic child to understand its meaning.

5.2 Future Work

Though the words in our simulation were already segmented, the real world provides no such easy segmentation. The child may assume that "I" is merely one syllable of another structure that has been repeated often. This appeared to be the case with the normal child in our script, who ritualistically said "I got it" no matter who caught the ball. The infant's mastery of the structure "I want" before the more general use of "I" has also been noted in the literature [5]. Autistic children may have more difficulty with this as well, as they are known to repeat sentences verbatim without making use of their constituent words [4].

To test these ideas, we plan to remove the context-free language model from our speech recognizer, and deal with raw phonemes instead. Though our oversimplification of speech-as-text is still relevant if children learn to segment pronouns before they learn what the words mean, we would like to make our language learning model far more general. At the early stages of word learning, segmentation may not even be complete before associations occur. Thus, an important next step for our model is to deal with this issue by working directly with the phonemes themselves.

Finally, this work is a milestone along the way to the larger goal of building a robot that has a grounded concept of self. By connecting this model to our earlier work on identifying the self in the visual image [26], we hope to make a robot that can learn, reason, and communicate about its own physical properties. In doing so, we hope to better understand how each of us comes to identify with the word "I."

6. REFERENCES

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

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

[3] E. V. Clark, "From gesture to word: on the natural history of deixis in language acquisition," in *Human growth and development: Wolfram College lectures 1976*, J. S. Bruner and A. Garton, Eds. Oxford: Oxford UP, 1978.

[4] 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.

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

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] S. R. H. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, 2000.

[14] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Sun Microsystems, Tech. Rep. TR-2004-139, Nov 2004.

[15] J. A. Anderson, *An Introduction to Neural Networks*. Cambridge, MA: MIT Press, 1995.

[16] G. Boole, "The laws of thought," in *George Boole's Collected Logical Works, Vol. 2*. La Salle, Illinois: The Open Court Publishing Company, 1854/1952.

[17] L. G. Valiant, *Circuits of the Mind*. New York: Oxford UP, 1994.

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

[19] J. N. B. III, "Transcript /eng-usa/bohannon/bax/leah1.cha," 1976.

[20] J. N. B. III and A. L. Marquis, "Children's control of adult speech," *Child Development*, vol. 48, pp. 1002–1008, 1977.

[21] E. L. Stine and J. N. B. III, "Imitations, interactions, and language acquisition," *Journal of Child Language*, vol. 10, pp. 589–603, 1983.

[22] B. MacWhinney, *The CHILDES project: Tools for analyzing talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[23] S. Baron-Cohen, *Mindblindness*. Cambridge, MA: MIT Press, 1995.

[24] G. Bradski, A. Kaehler, and V. Pisarevsky, "Learning-based computer vision with intel's open source computer vision library," *Intel Technology Journal*, vol. 9, no. 1, 2005, available online.

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

[26] P. Michel, K. Gold, and B. Scassellati, "Motion-based self-recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, 2004.