Using Speech and Natural Language Technology in Language Intervention

Jill Fain Lehman

March 19, 1997
CMU-CS-97-119

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

Educational and clinical techniques for language intervention in children with autistic spectrum disorders (ASD) focus on achieving a complete, speech-to-speech, communicative loop. To date, the AI technologies developed in areas like speech recognition, natural language processing, student modelling and intelligent tutoring have not been applied to the specific needs of children with ASD. In this paper we describe the design of *Simone Says*, a proposed software environment in which young children can practice semantically and socially meaningful language by playing a sort of interactive, linguistic game of *Simon Says*. Current research and practice in remediation both stress the need for achieving engagement and sustaining motivation in taking appropriate conversational turns and using language in functionally appropriate ways. *Simone Says* is intended to meet these requirements by using the natural attraction of computers to create opportunities for meaningful, speech-based language practice in a highly simplified social setting. In exercises that progress from vocabulary building to simple social conversation, the system will automatically generate contexts in which the student is rewarded for meaningful responses as defined by his or her current position along the normal developmental progression.

1. Motivation

The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) defines *pervasive developmental disorders* (alternatively, autistic spectrum disorders (ASD)) as a syndrome along three dimensions:

- Qualitative impairment in reciprocal social interaction,
- Qualitative impairment in communication, and
- Restricted, repetitive or stereotyped patterns of behavior, interests or activities [2].

Because the disorder is syndromic, subsets of symptoms and their severity vary across individuals, but onset in at least one area of dysfunction must occur before age three for this diagnosis. ASD is a neurologically-based, life-long disability occuring in about 2/1000 individuals. Among children the disorder is more common than either Down Syndrome or incidence of childhood cancer. With health care and education costs near $20,000 per year per child, a conservative estimate of disorder-related expenditures for children is $1.4 billion annually [10, 29].

Current clinical, social, and educational policy is designed to take advantage of critical periods in language development and neural plasticity by focusing on early detection and intervention. Although there has been extensive debate over which type of impairment constitutes the primary deficit of the disorder [4, 5, 9, 15, 19, 49, 50, 60, 62, 68], we cannot overestimate the importance of establishing a basic language capability in children with ASD. Research has shown that meaningful speech by school-age is the single most predictive element of a favorable long-term prognosis [17, 54, 57]. In more immediate terms, deficits in verbal expressive language have been found to be the most stressful type of impairment with which parents of children with ASD must cope [8]. Moreover, our ability to advance social and behavioral development may well hinge on improving the child’s ability to communicate [53].

Some children with ASD never progress beyond the most basic forms of non-verbal communication. Others speak, but remain predominantly echolalic — repeating the words and phrases of others with little or no understanding of the structure of language — well into their school-age years. Those who do eventually acquire functional language seem to do so in the normal progression, albeit with significant delays and some noticeable areas of underachievement [65, 66]. In particular, children with ASD invariably have trouble with the pragmatic aspects of language — when, how, and why language is used to achieve goals in interactions between people. Thus, the characteristic delays in the lexical, syntactic, and semantic levels of language development seem to stem from difficulties in understanding and constructing the pragmatic context in which normal acquisition occurs. One of the great developmental mysteries is how normally-developing children can acquire language simply by being in a linguistic community. The case of children with ASD suggests that the communicative function of language — the pragmatics of the discourse situation in which most children effortlessly exist — adds enormous constraint to the task of inducing the linguistic rules of their environment. Without that information, the ‘‘problem of language’’ is made more difficult or, for some, insurmountable.[1]

The history of applying technology to the communicative problems of ASD is brief. Colby had some initial success in using computers to instill an interest in speech-related sounds and language in mute children with autism in the early 1960's [14]. Since then, however, efforts have centered on providing augmentative technology (e.g. picture boards, communication devices) for children who remain essentially nonverbal. For those who show some verbal behavior (echolalic or productive), little that is specific to their problems has been done unless and until they begin reading [24, 51, 63]. The state of technology for language intervention defined more broadly includes many innovations, but little that addresses the needs of this population. Current software options consist primarily of comprehension drill, with interaction that is mouse- or keyboard-based rather than verbal. Software providing speech-based turn-taking targets only the acoustic level, with a focus on reinforcing prosodic and/or paralinguistic features such as pitch and duration [28, 30].

In contrast to the current focus of technology, educational and clinical techniques for stimulating language in children with ASD focus on achieving a complete, speech-to-speech, communicative loop. Regardless of whether the conversational context is essentially therapist-centered [40] or child-centered [18, 22, 31], research and practice both stress the need for achieving engagement and sustaining motivation in taking appropriate conversational turns and using language in functionally appropriate ways. We believe that current technology in speech, natural language processing, and graphics can help meet this need.

Two factors lead us to conclude that there is untapped potential in software that provides verbal turn-taking in a true communicative loop. The first factor is practical. Intensive one-on-one therapy as early in life as possible seems to be the treatment with the most efficacy [43]. Yet, it is unrealistic to expect that the majority of the families of young children with ASD can afford such treatment by professionals, or that family members have the time, energy and knowledge to act as effective paraprofessionals. In short, appropriate software can augment the resources demanded of families, schools, and society at large [25].

The second factor contributing to our conclusion is the inordinate interest in computers shown by many children with ASD [70]. Following the advice of Temple Grandin [21], a successful Ph. D. and entrepreneur who is also autistic, the idea is to use the naturally engaging power of computers to turn perseveration into progress. If full human-to-human communication is overwhelming or aversive for these children (particularly in instructional settings [56]), but human-computer interaction is manageable and attractive, then it seems appropriate to ask how we can use the medium to further our teaching goals. We take the view that computer-based interaction is a particular kind of *environmental engineering* [23, 69], one in which variability in

---

[1] The fact that a small percentage of children will eventually achieve functional language more or less on their own should not be interpreted as a counterargument to intervention. The significant language delays experienced by these children contribute to the stress of caring for a child with ASD, compound the social disadvantage inherent in the disorder, and put them at an educational disadvantage with respect to their peers. Intervention may not be the deciding factor in their eventual language competence, but it may help mitigate the associated problems, enabling more age-appropriate behavior. A recent study of ''autism in the third generation'' found that children born after 1974 tended to show better language and social skills at diagnosis than those born earlier, probably as a result of early intervention services in speech and language [16].

prosody, lexicalization, syntactic structure, semantics, and pragmatic context can be systematically controlled and the children's visual strengths exploited [13, 61].

Imagine a continuum with the total predictability of a much-loved video at one end and constant novelty of human-to-human communication on the other. The sort of human-computer interaction we propose involves principled movement along this line. The point of the proposed software is *not* to replace human interaction, but to help provide essential practice in language subskills. Technology can help to do this by providing a series of interactive experiences of increasing complexity at a rate that ensures that earlier stages of language development have become highly practiced and automatic before experiences based on later stages are presented. The assumption underlying our approach is that the skill automatization that results from practice in the simplified environment will, at each move along the continuum, help to reduce cognitive load enough to enable learning the next step. Cognitive and psycholinguistic theories that provide the rationale for this view of language include ActR [3], Gibson's work in memory limitations and language processing [20], and in particular, Soar [38, 39, 46, 47].[2]

2. Simone Says

In this section we describe a particular piece of software, *Simone Says*, its rationale, and the existing technologies that support its development. *Simone Says* is a sort of linguistic *Simon Says*, where Simone is a character (shown in Figure 2) that models appropriate language in the program's simple environment.

2.1. Design and Rationale

Simone Says is intended to create opportunities for meaningful language practice in a highly simplified social context. The purpose of the program is to ‘‘bootstrap,’’ or otherwise lead children through the normal developmental sequence, from Brown's Stage I until early Stage IV [11]. In general terms, the linguistic targets of the program are:

- A core vocabulary of 100-200 words
- Basic syntax and semantics over the core vocabulary
- Simple pragmatics and joint attention
- Conversational turn-taking
- Simple conversational repair

In a normally-developing population this would correspond to a portion of the acquisition that occurs between 18 and 36 months (i.e., mean length of utterance (MLU) from 1.0 to 3.9). Of course, in our target population it is much more likely that children falling in this range for MLU will be significantly older, approximately kindergarten age or above. In order to provide practice in language-specific skills and a closer approximation to a true communicative loop, interaction

[2]In addition to asking the question, *Can we help remediate the language-disordered behavior in ASD using a simplified social interaction?*, a second theoretical question to be addressed by this research is *Can echolalic behavior be harnessed?* In observing interactions between our system and echolalic children, we may find that providing them with very small, simple pieces to echo in appropriate pragmatic contexts facilitates their gradual shift from gestalt to analytic processing [52, 54].

with the system will be through speech rather than gesture. The initial versions of *Simone Says* will be appropriate for children who have already demonstrated at least minimal verbal communication, that is, children who vocalize at least one or two words reliably in appropriate contexts and who do not use those same words in inappropriate contexts. Thus, issues involved in moving children from the pre-linguistic to emerging language stage are beyond the scope of this research. Teaching pronunciation per se is, similarly, not our goal, although the technology we will use allows some flexibility in recognizing approximations to words.

The design of *Simone Says* is motivated largely by the need to teach the efficacy of language as a vehicle for making our thoughts and desires known to others. The system's basic interactive loop is shown in Figure 1. It consists of (1) the presentation of a visually-simple graphical stimulus, (2) the production of a referentially meaningful speech act by the child (or modelled by Simone or one of the other characters), and (3) a natural-consequence animation sequence as reward. In other words, each interaction directly reflects the idea that meaningful spoken language influences the behavior of others. All interactions with the program teach this lesson, whether they are simple one-word utterances or more complex utterances expressed within a simple conversational context.

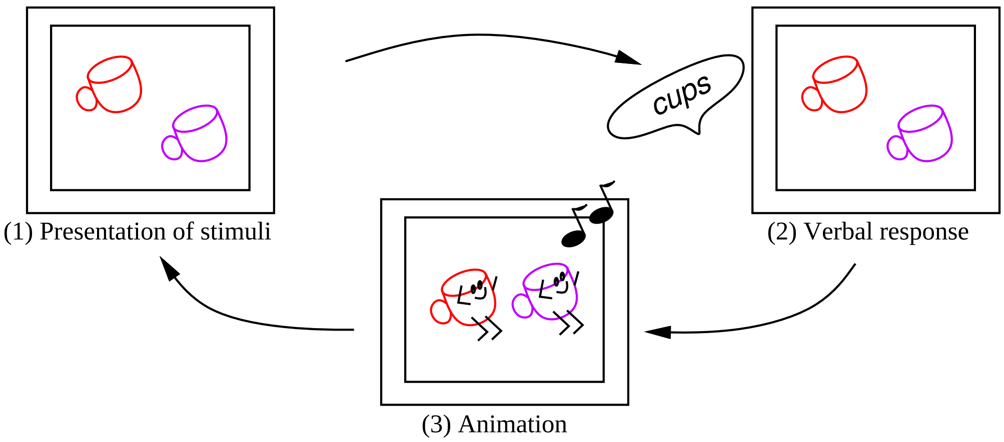

Figure 1: The basic interactive loop in *Simone Says*

Let's examine each aspect of the interaction in turn. The first step is the presentation of a visually engaging but graphically simple stimulus. The core vocabulary consists of common, everyday objects and actions, both to teach functionally useful vocabulary and to maximize the likelihood of practice and transfer in the home and school settings. Graphical simplicity is necessary both for computational reasons (the higher cost of animating a complex scene) and to help ameliorate problems with distraction and overspecificity in encoding that are characteristic of the disorder [32].[3] Although the stimuli are intended to be simple, multiple examples can be

[3]Unusual fears are characteristic of the disorder, as well, making it possible that some of the stimuli may cause anxiety in some children. To guard against triggering a fearful response, the graphics and animation databases will be parameterized so that particular items or sounds can be selectively disabled. A point-and-click interface will be created to allow parents, educators, and therapists to create an individualized profile of ‘‘safe’’ items prior to any interactions.

generated within relevant dimensions of variability (color, size, position in relation to other objects on the screen, background) in order to increase the likelihood of generalization. Figure 2 shows the interface to *Simone Says* with its four distinct screen areas: the control panel (upper left), the text echo box (upper right), and Simone's location (lower left) adjacent to the stimuli box (lower right). Presentation of each new stimuli follows the same pattern: a short animation sequence designed to focus the child's attention on the stimuli box, cessation of animation and movement by Simone to cue the response, followed by an individual-length response pause. For example, Figure 2 shows an example of the interface after a focus animation in which the apple comes onto the screen spinning, then slowly settles into its position.

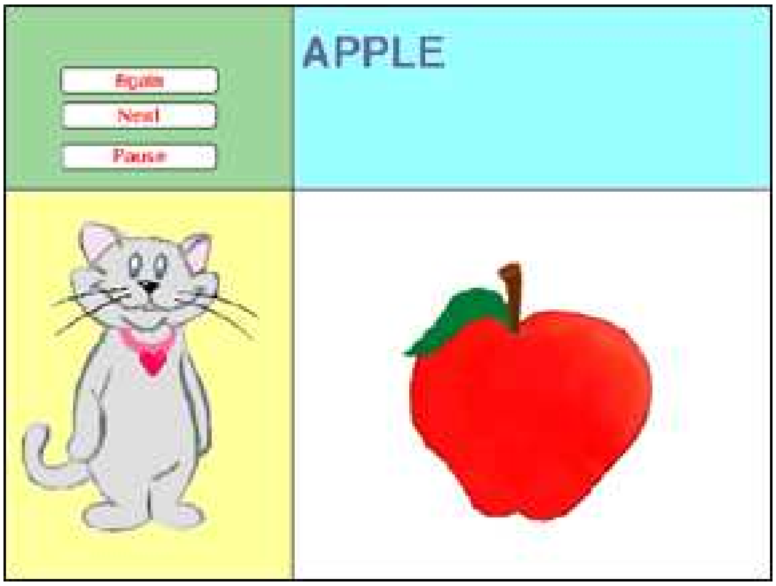

Figure 2: The interface consists of four distinct areas: a control panel (upper left), a text echo box (upper right), and Simone's location (lower left) adjacent to the box where stimuli are presented (lower right).

Producing a referentially meaningful response is the second step in the interactive loop. Note that the conversation is user-initiated (although admittedly within a rigidly defined context). In other words, it is the child that decides which object(s) to talk about from those visually available and what to make the object(s) do.[4] In all instances, only referentially meaningful utterances will produce a response, with Simone modelling an appropriate utterance if the child cannot produce one. Speech by either the child or a character is echoed in the text box, as shown in Figure 2. The system will automatically track the individual's history with the elements of the stimuli across linguistic targets. It will use this model of the child's current competencies to generate both examples that afford practice of acquired skills and those that require a skill that is slightly

[4]Current technology precludes a purely child-centered teaching style at the level of linguistic phenomena we are trying to support. In addition, almost any automated interaction based on teaching a particular developmental sequence will have some resemblance to a therapist-centered or discrete trials approach. *Simone Says* is an attempt to find a midpoint between child-centered and therapist-centered interaction, with rate of presentation, focus, and criteria for success under partial control of the child. As a method of language teaching, then, our approach falls somewhere between the natural and analog schools [18] with an emphasis on motivating early social communication [26, 41, 54]. In terms of developmental theory, the system embodies an *active person-active environment* design [58].

more difficult in the normal developmental progression.

The final step of the interaction is the reward of a natural-consequences graphical animation that reflects the child's utterance. For example, in Figure 2 the apple bobs and spins in response to being appropriately labelled. We are in regular contact with area therapists and speech/language pathologists to discuss stimulus design. In addition to their feedback, we believe three principles are important in designing the animation sequences:

1. **Make every interaction rewarding.** In other words, playing the game must itself be reinforcing [31]. For this reason, we choose action sequences that are particularly appealing to children with ASD (spinning, jumping, swinging, splashing, lining up) as well as include the sorts of exaggeration and slapstick amusing to most children.[5]. In addition, the ability of the system to always model some appropriate response for the child ensures that each interaction is a no-lose situation; some kind of animation always results.

2. **Motivate active involvement.** Because a character will always, eventually, produce an utterance that results in an animation, it is imperative that we construct the system to keep the child motivated to produce meaningful language rather than passively receive the reward by relying on Simone. Since predictability and control are enormously important to children with ASD, we assume that successfully making the system do what was intended by the child is intrinsically more rewarding than the less predictable response that comes from letting Simone choose the focus (i.e. presented with a stimuli as in the leftmost frame of Figure 4, Simone might choose to say ‘‘Jump’’ rather than ‘‘Eat’’). We can also take advantage of the inherent impatience of children, and increase the duration of the pause that occurs before modelling in relation to the degree of success the child has had with this sort of stimuli and task in the past.

3. **Balance realism with fun.** While it is generally accepted that natural consequences are more reinforcing and lead to better generalization, the notion of natural consequences in *Simone Says* is limited to making the action referentially connected to the scene. A referentially-connected reinforcer provides a natural consequence in the sense of demonstrating the efficacy of verbal language.[6] However, the notion should not be taken too far. Apples that can only be eaten are considerably less engaging than apples that can line themselves up, spin, dance or sing. This is not to imply that all objects will be able to perform all actions regardless of true semantic constraints. On the contrary, part of Simone's role as modeller will be to indicate when a semantic constraint has been violated and offer an appropriate alternative (‘‘Gee, trains can't drink, but they can move. Move train!’’). However, to keep engagement and enthusiasm high, it does seem useful to treat a few verbs as more generally applicable than they truly are.

Within the confines of this basic interactive loop, the child must be challenged to progress along the developmental dimensions of vocabulary, syntax, semantics, and pragmatics. The key

---

[5]The use of visuals normally associated with self-stimulating behavior may seem controversial, however, recent research has shown that using perseverative behaviors as reinforcers produces the best task performance, and in many cases actually decreases non-task-related perseveration [12]

[6]A disk that spins in the upper corner of the screen, a baseball player that advances around bases, and a bell that rings when a thermometer's mercury reaches the top are all examples of reinforcers that are *not* referentially meaningful for the stimuli of an apple but that are, nonetheless, typical of current software design.

to leading the child forward lies in slowly expanding the definition of what constitutes a referentially meaningful response, that is, by changing the criteria for success that triggers a rewarding animation. Figures 2 through 6 demonstrate this idea, showing how the same basic stimuli can be reused in increasingly complex contexts requiring increasingly complex language. Figure 2, as we've already seen, introduces the icon for *apple* while expecting only the simplest communicative act, labelling an object that is already a focus of attention. Once the child begins to show mastery of this task across a number of visually distinct episodes for a number of concrete nouns, the system might begin to introduce stimuli to teach the plural, as in the left frame of 3. In this situation an utterance of ''apple'' would produce only a simple animation of a single referent, reinforcing the meaning of the response (e.g. the single spinning apple in the middle frame of the figure). To lead the child to the next step, however, Simone would model ''apples,'' resulting in a more interesting animation involving all the relevant referents lining up and forming a train (the right frame of 3).[7]

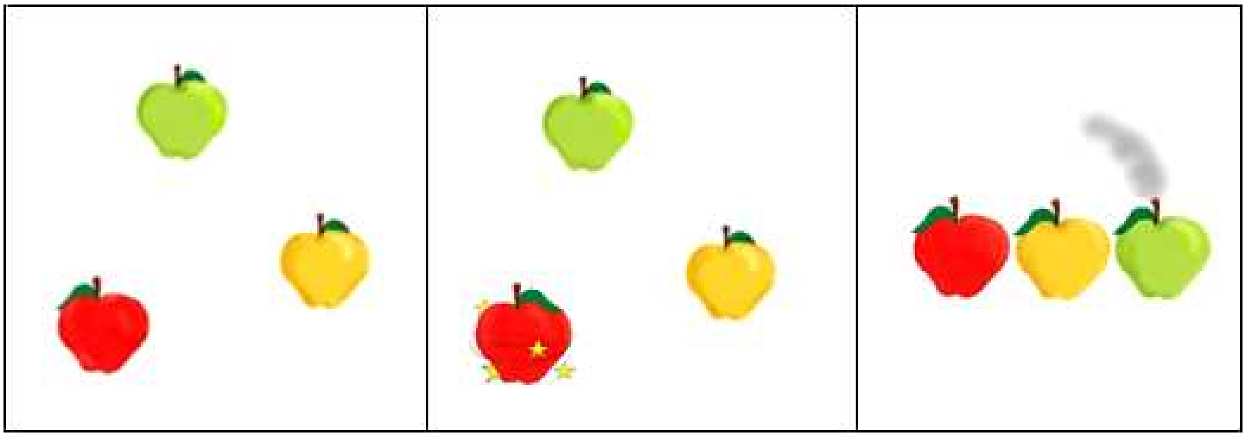

Figure 3: A sequence for introducing the plural form. The left frame shows the stimuli at the end of the focus animation. The middle frame shows the reward animation given for ''apple,'' while the right frame shows the richer reward for the plural.

As an alternative to introducing the plural morpheme, Figure 4 shows how the introduction of an actor into the scene during the focus animation provides the opportunity to model a more complex utterance along the vocabulary dimension (from concrete noun to verb) with ''eat.'' Later, essentially the same sequence can be used to move the child along the syntactic dimension by requiring both the noun and verb; ''eat apple'' or ''apple eat'' would be considered acceptable although either might occasion subsequent modelling of ''Yea! Eat the apple!'' by Simone.

Movement along the pragmatic dimension requires establishing joint reference with one of the animated characters, an extremely difficult task for children with ASD [42]. Figure 5 continues the linguistic progression for *apple* started in the previous figures. Here, the language already mastered is adequate to the task (''Eat the apple'' or ''Eat the banana'' being the simplest targets). A response of ''Eat'' alone fails to convey enough information to achieve the communicative goal, and should result in a simple subdialog (''Eat what?'' or ''Eat the apple?'')

---

[7]Simone is always on screen; in consideration of space, we omit all but the contents of the stimuli box in this and the remaining figures.

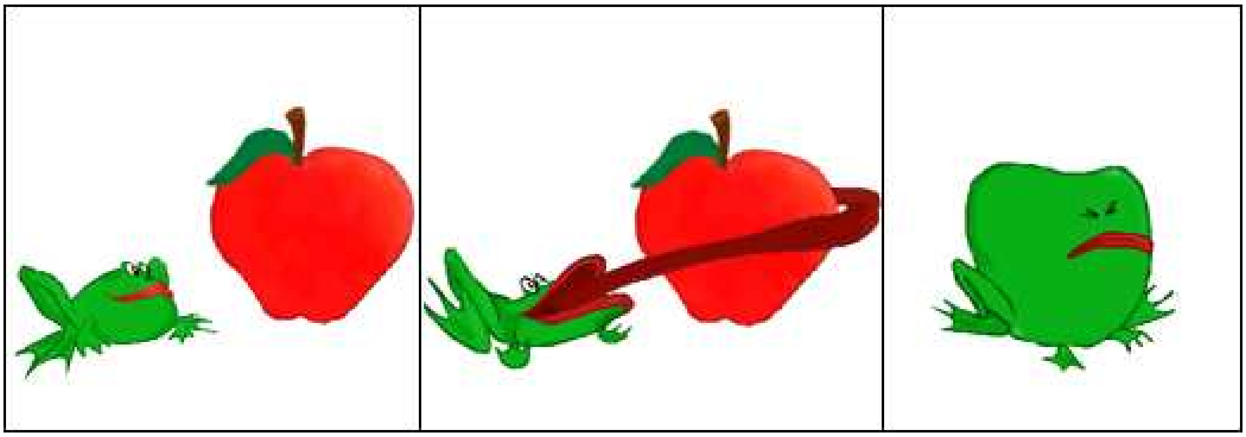

Figure 4: A sequence for introducing the verb ‘‘eat’’ and noun-verb combinations.

with Annie (the frog) or Simone. By varying the object to be chosen along relevant dimensions — e.g., two apples of different colors, two doors of different sizes, a book on a table versus one that is under the table — scenes like this teach how perceptually available features can be used to disambiguate reference.

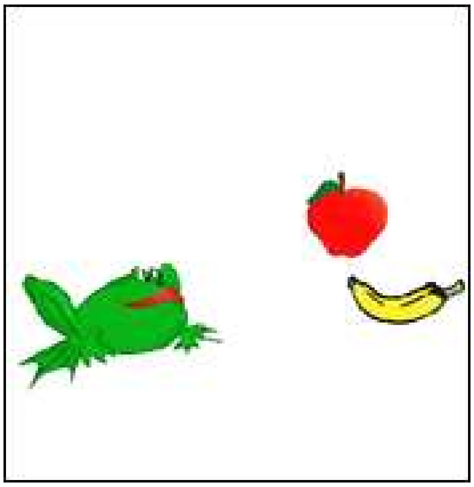

Figure 5: A sequence for teaching how perceptual features can disambiguate reference. A response of ‘‘Eat’’ alone is inadequate, an object for the verb is required.

Figure 6 goes a step further by embedding language in a simple social context. Following Prizant and others [6, 7, 48, 53], we explicitly — and visually — model for the child the connection between mental state and communication. As shown, we accomplish this by using a thought bubble with a miniature version of the target animation played inside it as a second response cue. The point is to make explicit the link between the intention to produce an action and the language that makes that intention known to others. If this second sort of cueing still does not produce an appropriate response, then the characters involved might cue with a question, or simply model the response.

Situations like the one shown in Figure 6 tax the ability of the technology to anticipate the child’s responses with reasonable accuracy and, thus, represent the most sophisticated sort of communicative interaction we will provide. These simple social situations allow us to introduce

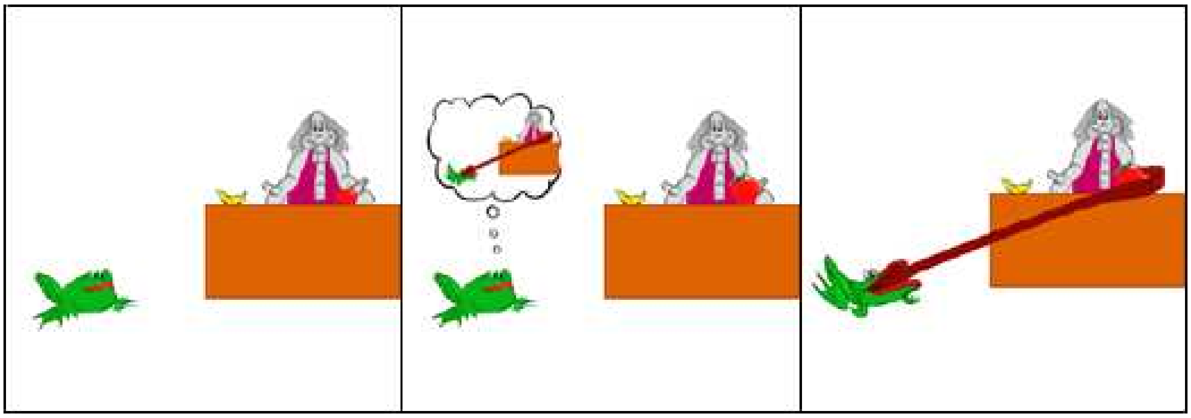

Figure 6: Embedding language in a social context and modelling *theory-of-mind*.

short verbal scripts and can be used and reused to target a variety of pragmatic issues, such as point of view (‘‘Take the apple’’ versus ‘‘Give the apple’’) and wh-questions, all of which may call for more complex turn-taking behavior as well as elementary conversational repair.

2.2. Supporting Technology

Arguing for the efficacy of computer technology in language intervention is not a guarantee that the technology itself is up to the task. In this section we discuss the uses of and problems with current technology in terms of the subtasks for *Simone Says*. Figure 7 shows the basic processing loop (in boldface) and knowledge bases (in italics) required for the interaction pictured in Figure 1.

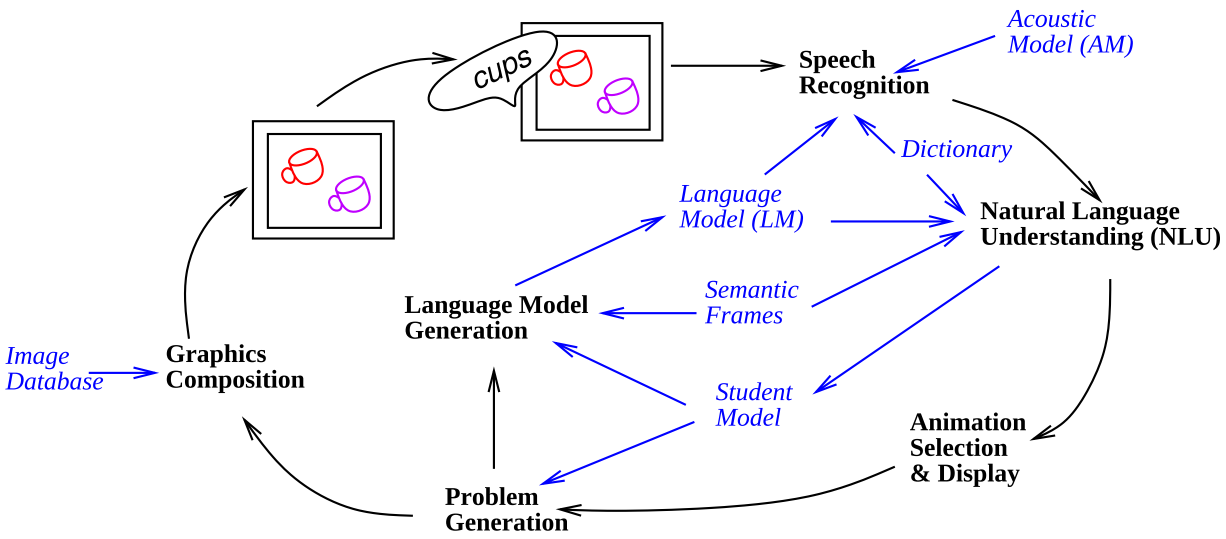

Figure 7: Processing loop and knowledge bases for *Simone Says*

We begin at the top of the figure with the voice input to a speech recognizer. Current speech recognition programs for continuous, speaker-independent, large vocabulary domains are available both commercially (e.g., Microsoft’s Whisper or AT&T’s Watson) and from university research labs (e.g. SPHINX [27]). As shown, the knowledge bases used by the recognizer are the Acoustic Model (AM), which transforms the speech waveform into phonemes, and the Language Model (LM), which maps phonemes into the morphemes in the dictionary. Acoustic models for existing systems have generally been trained on very large corpora from adult speakers. Since the pronunciation and vocal characteristics of adults differ significantly from those of young children, these acoustic models will have to be adapted to our target population. Adaptation requires only a relatively small corpus of speech to be collected (as outlined in Section 3). Articulation problems and phonemic confusions beyond what is normal for chronological age are not a characteristic of verbal children with ASD [66]; the sorts of prosodic differences that are characteristic [60] are filtered out during the initial phases of speech processing. As a result we may be able to improve the accuracy of the adapted AM by combining the modest corpus we intend to collect with one of slightly older children that has already been collected [45].

Despite their ability to handle vocabularies of 50,000 words or more, speech systems nevertheless impose significant limits on the complexity of the grammar they can recognize. The point of this proposal is *not* to conduct basic research in speech technology but rather to use the technology that exists in a new and clinically-informed way. In *Simone Says* the usual restrictions on the complexity of the Language Model are unlikely to have an impact on accuracy for a number of reasons. First, the target vocabulary itself is quite small: probably less than 3000 morphemes. Second, and most important, is the constraint that comes from having total control over the stimuli; since we define what is referentially meaningful for each example, we believe we can generate the appropriate LM on an example-by-example basis.[8] Within these constraints it seems likely that current technology can support the sort of very limited mixed-initiative interaction we envision, although this is clearly an empirical question.

While accuracy of the recognizer is critical to keeping the rate of rejected utterances low, it is unreasonable to expect perfect recognition. Thus, the first task of the Natural Language Understanding (NLU) component is to compensate for misrecognitions on the part of the recognizer. The accurate recovery of every morpheme is not necessary; some may, in fact, be irrelevant or redundant. However, those morphemes that carry the meaning of the utterance must be recovered so that the student's progress can be charted and the appropriate animation selected. Ameliorating this problem is the fact that out-of-vocabulary words, a typical source of misrecognitions, are unlikely in this task with this user group.

The second function of the NLU component is to recognize both positive changes and errors in the student's constructions. The NLU system we intend to use as the basis of this component is CHAMP, a system originally designed to learn user-specific grammars through interactions with a user performing a routine task [1, 33, 34, 35, 36, 37, 64]. Starting with a small core grammar and semantic representation for the task domain, CHAMP understands each utterance typed by

---

[8]The LM for a given scene is a function of the stimuli and the student's proximal zone of development [67], as defined by the student model. It can, we believe, be constructed using the referentially meaningful utterances of length less than n composed from vocabulary defined for the stimuli. The value of n and set of meaningful utterances varies over time; at the beginning of the one-word stage we might expect that a picture of an apple would be meaningfully referenced only by ‘‘apple,’’ or ‘‘fruit.’’ Later, as the student begins to show mastery of verbs, we might expect ‘‘eat,’’ still later ‘‘eat apple.’’ Of course, this approach is computationally feasable only because, between Stages I and IV, n always remains very small.

the user as more or less deviant with respect to that grammar. Deviant utterances cause the creation of new grammatical elements so that the user's particular, often idiosyncratic, grammar can be understood efficiently in future interactions. It is easy to see how this capability can be used in *Simone Says*. For each stimuli, the expected Language Model defines the core grammar, and can be constructed on the basis of the student's current strengths and potential next steps. CHAMP then views the child's utterance in terms of this LM, pinpointing sources of deviation. Utterances that are non-deviant represent growing mastery and advances along the developmental continuum. Deviations that cannot be corrected by assuming next-best guesses from the speech recognizer can be attributed to the user and form the basis for updating the student model and choosing the next example.

Once NLU has assigned a meaning to the utterance in terms of its library of semantic frames, that representation can be used to choose the appropriate animation sequence. There are a number of commercially-available packages for authoring 2D animations on PC and Macintosh platforms that are more than adequate for the kinds of scenes in *Simone Says* (the figures in the previous section are taken from animations created using Macromedia's Director5). If the child's response has been inappropriate (or has not been forthcoming), the animation must include modelling by one of the characters that inhabits this simple social world. The system's ability to focus remediation on specific errors will depend on the accuracy of the recognition and understanding process; it is more confusing to pinpoint an error incorrectly than to simply have Simone model something referentially appropriate.

While the user's attention is held by the animation, the system must do the processing required to generate the next example. This process is based on the updated student model provided by CHAMP. The student model is the structure that ties together the three types of processes in *Simone Says*: language, problem generation, and animation. The model both records the functionally useful responses for each kind of stimuli (to track generalization) and specifies the uneven border that constitutes the child's developing language (he or she may, for example, still be acquiring words for some stimuli but combining words for others). As Figure 7 shows, problem generation feeds into both the component that generates the Language Model for the new example and the component that produces the new graphical image. Once these two structures have been created, the basic interactive loop can begin again.

2.3. Evaluation

The ultimate goal of *Simone Says* is, of course, to help children with ASD acquire functionally useful language. Thus, evaluation of the program will be oriented to answering the following questions:

1. Is there demonstrable growth in language during human-computer interaction as measured by (a) increased number of appropriate responses, (b) increased complexity of responses as measured by MLU, (c) decreased latency of response, (d) decreased amount of response modelling, and (e) generalization of response across stimuli?

2. Is there demonstrable growth in language during human-human interaction, as measured by appropriateness and complexity of response?

3. Can any such growth be attributed in part to the software intervention?

Because we are interested in tracking changes across the developmental progression, our intent is to conduct a longitudinal study of verbal children with ASD using *Simone Says* for about one year. Transition from Stage I to Stage IV generally takes 18 months in normally developing children, longer in children with ASD. However, not all children will begin at the same stage in our study. As long as we have a reasonable number of children starting at each stage, we should be able to see some evidence for efficacy across the various linguistic targets within a year.

Answering the first question posed above is straightforward since the measures involved can be collected automatically as part of building the student model. Answering the second question requires some interval-based assessment in the home or school setting. We intend to collect language samples via videotape three times, at the beginning of the study, at six months, and at the end. Transcriptions of the video will be scored using the Index of Productive Syntax (IPSyn, [59]) and Prutting and Kirchner's Pragmatic Protocol [55] or similar instruments.

To answer the third question we will use a standard experimental vs control design, with half our subjects receiving intervention with *Simone Says*, and half receiving no software intervention. The dependent variables will include the IPSyn and pragmatic scores, but we do not expect these scores alone to be revealing. The children involved in our study will undoubtedly be participating in a variety of other therapies at the same time, many more frequent and intensive than exposure to *Simone Says*. Moreover, since we are choosing the stimuli specifically to afford transfer in everyday situations, we expect children in both conditions to advance linguistically. With so many possible sources of language remediation, we do not expect gross-interval measures to show large differences between the conditions. Moreover, a lack of significant difference between groups would not necessarily be evidence that *Simone Says* is ineffective. Our point is not to prove that children *must* use our software to progress, but to explore whether *Simone Says* contributes effectively to that growth. In other words, although showing that the software can significantly speed up language development would be highly desirable, it is an equally useful outcome to show that we can sustain the rates available through current levels of human intervention at a lower cost.

In order to assess whether *Simone* is making a contribution, then, we need a finer-grained evaluation than the three-time videotape record. The exaggerated level of encoding specificity in children with ASD combined with simple practice effects predicts significant differences on trained versus untrained items, at least in the short run. Thus checklists of the items in the full stimuli set will be provided to the home and school of each child to chart shifts in usage on a weekly or monthly basis (the IPSyn and Pragmatic Protocol collection can serve as a check on the accuracy of these reports). If *Simone* is useful, we would expect a different acquisition profile for the two conditions, with an increased likelihood for trained items to appear in at-home vocabulary in the experimental condition. Effectiveness in the natural environment can be claimed unambiguously if there is differential improvement in the trained items for the experimental group, even though such differences may be transient as the influences of other linguistic experiences accumulate.

In ideal circumstances, the outline for evaluation given above would be extended to include daily use of the software by providing training and hardware to families willing to participate. This would allow us to guage frequency-of-use as a factor if no difference in performance between the groups in the less-intensive conditions were found.

3. Feasibility and Project Plan

Simone Says is an ambitious project that relies on bringing together and expanding existing technology in new ways. Identifying and testing underlying assumptions early in the research is prudent because, although cheap to reproduce, sophisticated, robust software systems are expensive and time-consuming to build. We have identified three important issues regarding the feasibility of the system:

1. Acceptable accuracy in speech understanding for the target population: we are assuming both the ability to adapt an adult Acoustic Model and to generate Language Models on-the-fly.

2. Adequate commonality in reinforcers across users: based on feedback from experts in the field, we are assuming that there is a small set of types of animation that this heterogeneous community will find engaging.

3. Ability of users to tolerate the technology: we are assuming that use of a close-talk microphone, or, if necessary, a label microphone will not be aversive.

We believe that the best method for testing these assumptions is by an initial Wizard-of-Oz experiment (so-called because the user is expected to "pay no attention to the man behind the curtain"). In this type of experiment, a mock *Simone Says* is constructed with a human experimenter in the loop. The child receives the same type of visual stimuli and interacts via speech, as with the real system, but the interpretation of the response and selection of animation sequence is done in real-time by the experimenter. Although it requires a fair amount of training to ensure that the experimenter acts consistently and without undue intelligence, this method is a relatively inexpensive way to test our assumptions. During the experiment itself we would be able to see if the children could work with the technology and if our initial guesses about engaging animation were accurate. After the experiment, a portion of the collected speech samples would be used to adapt the Acoustic Model of the recognizer. Then the examples used in the experiment and the utterances given in response would be run through a skeleton system consisting of the recognizer, NLU component and Language Model generator to see whether an acceptable level of accuracy can be achieved.

In addition to testing assumptions, a Wizard-of-Oz experiment allows us to collect critical information for making informed design decisions once feasibility has been demonstrated. Since building interactive software is always an iterative process, data collected during the experiment can make an enormous difference both in terms of the number of iterations and the time to produce any particular version. The clinical and educational communities have expressed an interest and willingness to participate in this iterative process.

Note that the purpose of the Wizard-of-Oz experiment is to prove the feasibility of *Simone Says* as a piece of technology, not to prove the efficacy of the intervention it delivers. Consequently, the data collection period need not be as long as that outlined in the previous section. We believe feasibility can be demonstrated during the first phase of the following full project schedule:

- Phase I of System Building and Feasibility (18 months)
    - (6 months) Creation of graphical stimuli and software for data collection, feedback from professionals, experimenter training, and attaining consent of families.
    - (6 months) Data collection, extensions to speech recognizer and CHAMP.

- (6 months) data analysis and dissemination, Acoustic Model adaptation, and testing of the recognition-NLU-Language Modeling subsystem.

- Phase II of System Building (12 months)
    - Creation of remaining components (animation selection, problem generator and graphics composition), expansion of knowledge bases, and iterative feedback from professionals.

- Longitudinal Evaluation (18 months)
    - Experimenter training, attaining consent, at-home assessments, on-going experiment and data analysis.

4. Conclusions

Simone Says is intended to provide speech-based, functionally-oriented interactions for teaching language to children with ASD. The system will automatically generate contexts in which the student is rewarded for referentially appropriate responses as defined by his or her current position along the normal developmental sequence. The program will incorporate random variation in visual features to promote generalization, as well as automatic record keeping for charting progress.

To achieve this goal, there are three basic technical issues to be resolved: adaptation of current speech technology to the population, extension of current adaptive parsing technology to work with structures required by the speech recognizer, and creation of an underlying representation (the student model) that can be used to effectively coordinate speech, natural language, problem generation, and animation processes. The tools available for addressing these issues include mature speech and NL technologies from the research community and off-the-shelf authoring environments for creating animations of the quality found in commercial educational software. The main challenge, of course, lies in bringing together pieces with such different origins.

On the way to solving the technical issues we anticipate the creation of intermediate results that will be useful to other researchers in the fields of autism, computational linguistics, and language education. In particular, we will produce and make available both a corpus of child speech data and a database of the developmental sequences of 10 or more children with ASD. The former increases the amount of data available for adapting acoustic models in developing other speech-based software for children. The latter provides a longitudinal record of language change for a significant number of children against which other hypotheses can be tested.

By providing interactive experiences that range linguistically from vocabulary-building to simple social discourse, *Simone Says* will be the first software to create an environment where children can practice semantically and socially meaningful verbal language. As such, it represents the potential addition of an effective, low-cost option to the current intervention arsenal as well as a platform for exploring speech-based applications for the 3-5% of *all* children who enter school with a language disorder [44].

References

[1] Allen, C. S. & Bryant, B. R.
Learning a user's linguistic style: Using an adaptive parser to automatically customize a unification-based natural language grammar.
In *Proceedings of the Fifth International Conference on User Modeling.* 1996.

[2] American Psychiatric Association.
Diagnostic and Statistical Manual of Mental Disorders (4th Edition).
American Psychiatric Association, Washington, D. C., 1994.

[3] Anderson, J. R.
Rules of the Mind.
Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

[4] Ayres, J. A.
Sensory Integration and the Child.
Western Psychological Services, Los Angeles, CA, 1979.

[5] Baron-Cohen, S.
Social and pragmatic deficits in autism: Cognitive or affective?
Journal of Autism and Developmental Disorders 18(3):379-402, 1988.

[6] Baron-Cohen, S.
The autistic child's theory of mind: A case of specific developmental delay.
Journal of Child Psychology and Psychiatry 30:285-297, 1989.

[7] Baron-Cohen, S., Leslie, A. M., & Frith, U.
Does the autistic child have a *theory of mind*?
Cognition 21(1):37-46, 1985.

[8] Bebko, J. M., Konstantareas, M., & Springer, J.
Parent and professsional evaluations of family stress associated with characteristics of autism.
Journal of Autism and Developmental Disorders 17(4):565-576, 1987.

[9] Berkell, D. E.
Autism: Identification, Education, and Treatment.
Laurence Earlbaum Associates, Hillsdale, NJ, 1992.

[10] Bristol, M. M., Cohen, D. J., Costello, E. J., Denckla, M., Eckberg, T. J., Kallen, R., Kraemer, H. C., Lord, C., Maurer, R., McIlvane, W. J., Minshew N., Sigman, M., & Spence, M. A.
State of the Science in Autism: Report to the National Institutes of Health.
Journal of Autism and Developmental Disorders 26(2):121-154, 1996.

[11] Brown, R.
A First Language, the Early Stages.
Harvard University Press, Cambridge, MA, 1973.

[12] Charlop-Christy, M. H. & Haymes, L. K.
Using obsessions as reinforcers with and without mild reductive procedures to decrease inappropriate behaviors of children with autism.
Journal of Autism and Developmental Disorders 26(5):527-548, 1996.

[13] Charlop, M. H. & Milstein, J. P.
Teaching autistic children conversational speech using video modeling.
Journal of Applied Behavior Analysis 22(3):275-285, 1989.

[14] Colby, K. M.
The rationale for computer-based treatment of language difficulties in nonspeaking autistic children.
Journal of Autism and Childhood Schizophrenia 3:254-260, 1973.

[15] Dawson, G.
Autism: Nature, Diagnosis, and Treatment.
Guilford Press, 1989.

[16] Eaves, L. C. & Ho, H. H.
Brief report: Stability and change in cognitive and behavioral characteristics of autism through childhood.
Journal of Autism and Developmental Disorders 26(5):557-569, 1996.

[17] Eisenberg, L.
The autistic child in adolescence.
American Journal of Psychiatry 112:607-612, 1956.

[18] Elliott, R. O., Hall, K., & Soper, H. V.
Analog language teaching versus natural language teaching: Generalization and retention of language learning for adults with autism and mental retardation.
Journal of Autism and Developmental Disorders 21(4):433-447, 1991.

[19] Fay, W. H., and Schuler, A. L.
Emerging Language in Autistic Children.
University Park Press, Baltimore, MD, 1980.

[20] Gibson, E.
Memory capacity and sentence processing.
In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. 1990.

[21] Grandin, T., and Scariano, M. M.
Emergence: Labelled Autistic.
Arena, Novato, CA, 1986.

[22] Greenspan, S. I.
Reconsidering the diagnosis and treatment of very young children with autistic spectrum or pervasive developmental disorders .
Zero to Three (Bulletin of the National Center for Clinical Infant Programs) 13(2):1-9, 1992.

[23] Halle, J. W.
Arranging the natural environment to occasion language: Giving severely language-delayed children reasons to communicate.
Seminars in Speech and Language 5(3):185-197, 1984.

[24] Heimann, M., Nelson, K. E., Tjus, T., & Gillberg, C.
Increasing reading and communication skills in children with autism through an interactive multimedia computer program.
Journal of Autism and Developmental Disorders 25(5):459-480, 1995.

[25] Howard, J. R., Busch, J. C., Watson, J. A., & Shade, D. D.
The change-over to computer-based technology in early childhood special education.
Journal of Research on Computing in Education 23:530-544, 1991.

[26] Howlin, P.
Changing approaches to communication training with autistic children.
British Journal of Disorders of Communication 24:151-168, 1989.

[27] Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., & Rosenfeld, R.
The SPHINX-II speech recognition system: An overview.
Computer Speech and Language 7(2):137-148, 1993.

[28] IBM.
IBM Independence Series Speechviewer III.
IBM, NY, NY, 1996.

[29] Jensen, P.
Prevalence of autism and co-occuring disorders.
In *The Child With Special Needs Preconference on Autism, Washington D.C.*, pages handout. 1996.

[30] Kewley-Port, D., Watson, C. S., Elbert, M. Maki, D., & Reed, D.
The Indiana Speech Training Aid (ISTRA) II: Training curriculum and selected case studies.
Clinical Linguistics and Phonetics 5(1):13-38, 1991.

[31] Koegel, R. L, and Johnson, J.
Motivating Language Use in Autistic Children.
In Geraldine Dawson (editor), *Autism: Nature, Diagnosis, and Treatment*, pages 310-325. Guilford Press, 1989.

[32] Koegel, R., Egel, A., & Dunlop, G.
Learning Characteristics of Autistic Children.
In W. Sailor, B. Wilcox, & L. Brown (editor), *Methods of Instruction for Severely Handicapped Students*. Paul H. Brookes, 1980.

[33] Lehman, J. F., and Carbonell, J. G.
Learning the Users Language: A Step Towards Automated Creation of User Models.
In Wahlster, W., and Kobsa, A. (editors), *User Modelling in Dialog Systems*. Springer-Verlag, 1989.

[34] Lehman, J. Fain.
Supporting Linguistic Consistency and Idiosyncracy with an Adaptive Interface Design.
In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. 1990.

[35] Lehman, J. Fain.
Adaptive Parsing: A General Method for Learning Idiosyncratic Grammars.
In *Proceedings of the Sixth International Conference on Machine Learning*. 1990.

[36] Lehman, J. Fain.
Adaptive Parsing: Self-extending Natural Language Interfaces.
Kluwer Academic Publishers, Norwell, MA, 1992.

[37] Lehman, J. Fain.
Three Uses of Adaptive Parsing in Intelligent Tutoring.
In Engel, F. L, Bouwhuis, D. G., Bosser, T., and dYdewalle, G. (editors), *Cognitive Modelling and Interactive Environments in Language Learning NATO ASI Series*. Springer-Verlag, 1992.

[38] Lehman, J. F., Laird, J. E., and Rosenbloom, P. S.
A Gentle Introduction to Soar, an Architecture for Human Cognition.
In S. Sternberg and D. Scarborough (editors), *Invitation to Cognitive Science: Methods Models and Conceptual Issues*. MIT Press, 1997.

[39] Lehman, J. Fain, Newell, A. N., Polk, T., and Lewis, R. L.
The Role of Language in Cognition.
In Harman, G. (editors), *Conceptions of the Human Mind*. Lawrence Erlbaum Associates, Inc., 1993.

[40] Lovaas, O. I.
Behavioral treatment and normal education and intellectual functioning in young autistic children.
Journal of Consulting and Clinical Psychology 55:3-4, 1987.

[41] Loveland, K. A. & Landry, S. H.
Joint attention and language in autism and developmental language delay.
Journal of Autism and Developmental Disorders 16(3):335-349, 1986.

[42] McArthur, D. & Adamson, L. B.
Joint attention in preverbal children: Autism and Developmental Language Disorder.
Journal of Autism and Developmental Disorders 26(5):481-496, 1996.

[43] McEachin, J. J., Smith, T., & Lovaas, O. I.
Long-term outcome for children with autism who received early intensive behavioral treatment.
American Journal on Mental Retardation 97(4):359-372, 1993.

[44] Moore, M. T., Strang, E. W., Schwartz, M., & Braddock, M.
Patterns in Special Education Service Delivery and Cost.
Technical Report, Decision Resources Corp., Washington, D. C., 1988.

[45] Mostow, J., Roth S. F., Hauptmann, A. G., and Kane, M.
A prototype reading coach that listens.
In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 785-792. 1994.

[46] Nelson, G., Lehman, J. F., John, B.
Integrating Cognitive Capabilities in a Real-Time Task.
In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 658-663. 1994.

[47] Newell, A.
Unified Theories of Cognition.
Harvard University Press, Cambridge, Massachusetts, 1990.

[48] Ozonoff, S. & Miller, J. N.
Teaching Theory of Mind: A new approach to social skills training for individuals with autism.
Journal of Autism and Developmental Disorders 25(4):415-433, 1995.

[49] Ozonoff, S., Pennington, B. F., & Rogers, S. J.
Executive function deficits in high-functioning autistic individuals: Relationship to Theory of Mind.
Journal of Child Psychology and Psychiatry 32(7):1081-1105, 1991.

[50] Ozonoff, S., Rogers, S. J., & Pennington, B. F.
Asperger's syndrome: Evidence of an empirical distinction from high-functioning autism.
Journal of Child Psychology and Psychiatry 32(7):1107-1122, 1991.

[51] Panyan, M. V.
Computer technology for autistic children.
Journal of Autism and Developmental Disorders 14:375-382, 1984.

[52] Prizant, B. M.
Gestalt language and gestalt processing in autism.
Topics in Language Disorders 3:16-23, 1982.

[53] Prizant, B. M. & Wetherby, A. M.
Communicative intent: A framework for understanding social-communicative behavior in autism.
Journal of the American Academy of Child Psychiatry 26:472-479, 1987.

[54] Prizant, B. M., & Wetherby, A. M.
Enhancing Language and Communication in Autism: From Theory to Practice.
In Geraldine Dawson (editor), *Autism: Nature, Diagnosis, and Treatment*, pages 282-309. Guilford Press, 1989.

[55] Prutting, C., and Kirchner, D.
Applied Pragmatics.
In T. M. Gallagher and C. A. Prutting (editors), *Pragmatic Assessment and Intervention Issues in Language*. College-Hill Press, 1983.

[56] Romanczyk, R. G., Ekdahl, M., & Lockshin, S. B.
Perspectives on Research in Autism: Current Trends and Future Directions.
In Berkell, D. E. (editor), *Autism: Identification, Education, and Treatment*. Laurence Earlbaum Associates, Hillsdale, NJ, 1992.

[57] Rutter, M., Greenfield, D., & Lockyer, L.
A five to fifteen year follow up study of infantile psychosis: II. Social and behavioral outcome.
British Journal of Psychiatry 113:1183-1199, 1967.

[58] Sameroff, A.
The Social Context of Development.
In N. Eisenberg (editor), *Contemporary Topics in Developmental Psychology*. Wiley, New York, NY, 1987.

[59] Scarborough, H.
Index of productive syntax.
Applied Psycholinguistics 11:1-22, 1990.

[60] Schopler, E., and Mesibov, G.
Communication Problems in Autism.
Plenum Press, New York, NY, 1985.

[61] Secan, K. E., Egel, A. L., & Tilley, C. S.
Aquisition, generalization, and maintenance of question-answering skills in autistic children.
Journal of Applied Behavior Analysis 22(2):181-196, 1989.

[62] Siegel, B.
The World of the Autistic Child: Understanding and Treating Autistic Spectrum Disorders.
Oxford University Press, Oxford, England, 1996.

[63] Steiner., S. & Larson, V.
Integrating Microcomputers into Language Intervention.
Topics in Language Disorders 11:18-30, 1991.

[64] Sugar, D.
An Adaptive Medical Natural Language Parser.
Technical Report, Masters Thesis, University of Waterloo, 1994.

[65] Tager-Flusberg H.
A Psycholinguistic Perspective on Language Development in the Autistic Child.
In Geraldine Dawson (editor), *Autism: Nature, Diagnosis, and Treatment*, pages 92-115. Guilford Press, 1989.

[66] Tager-Flusberg, H., Calkins, S., Nolin, T., Baumberger, T., Anderson, M., & Chadwick-Dias, A.
A longitudinal study of language acquisition in autistic and Down syndrome children.
Journal of Autism and Developmental Disorders 20(1):1-21, 1990.

[67] Vygotsky, L.
Mind in society: The development of higher psychological processes.
Harvard University Press, Cambridge, Mass, 1978.

[68] Waterhouse, L., Fein, D., & Modahl, C.
Neurofunctional mechanisms in autism.
Psychological Review 103(3):457-489, 1996.

[69] Wetherby, A. M. & Prizant B. M.
Facilitating Language and Communication Development in Autism: Assessment and Intervention Guidelines.
In Dianne E. Berkell (editor), *Autism: Identification, Education, and Treatment*, pages 107-134. Lawrence Erlbaum Associates, 1992.

[70] Wilson, M. S.
Sequential Software for Language Intervention and Development.
Laureate Learning Systems, Inc., Winooski, VT, 1996.

Table of Contents

1. Motivation	**1**
2. Simone Says	**3**
2.1. Design and Rationale	**3**
2.2. Supporting Technology	**9**
2.3. Evaluation	**11**
3. Feasibility and Project Plan	**13**
4. Conclusions	**14**

List of Figures

Figure 1: **The basic interactive loop in *Simone Says*** **4**
Figure 2: **The interface consists of four distinct areas: a control panel (upper left), a text echo box (upper right), and Simone's location (lower left) adjacent to the box where stimuli are presented (lower right).** **5**
Figure 3: **A sequence for introducing the plural form. The left frame shows the stimuli at the end of the focus animation. The middle frame shows the reward animation given for ''apple,'' while the right frame shows the richer reward for the plural.** **7**
Figure 4: **A sequence for introducing the verb ''eat'' and noun-verb combinations.** **8**
Figure 5: **A sequence for teaching how perceptual features can disambiguate reference. A response of ''Eat'' alone is inadequate, an object for the verb is required.** **8**
Figure 6: **Embedding language in a social context and modelling *theory-of-mind*.** **9**
Figure 7: **Processing loop and knowledge bases for *Simone Says*** **9**