Toward the Use of Speech and Natural Language Technology in Intervention for a Language-disordered Population

Jill Fain Lehman
School of Computer Science
Carnegie Mellon University

Abstract

We describe the design of *Simone Says* an interactive software environment for language remediation that brings together research in speech recognition, natural language processing and computer-aided instruction. The underlying technology for the implementation and the system's eventual evaluation are also discussed.

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 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].

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 (see, e.g., [23]), 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 [21]. In day-to-day 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 [4]. Finally, our ability to advance social/behavioral development may well hinge on improving the child's communication.

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 [25]. 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.

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 [7]. Since then, however, efforts have centered on providing augmentative technology for children who remain essentially nonverbal. Some work has been done with modeling via videodisc [6], but for those who show some verbal behavior (echolalic or productive), little in the way of interactive software that is specific to their language problems has been available unless and until they begin reading [8, 24]. 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.

ASSETS 98 Marina del Rey CA USA

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 (e.g. [14]) or child-centered (e.g. [22]), 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, user modelling, and animation 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 and small group therapy as early in life as possible seems to be the treatment with the most efficacy [17, 22]. 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.

The second factor contributing to our conclusion is that the significant language delays and impairments in this population often co-occur with a marked preference for computer rather than human interaction [27], strong visual processing skills and rote memory, age-appropriate articulation, and preferential attention to language that is patiently repeated with little or no variation in prosody, word choice, or syntactic structure (for example, television commercials and videos). In other words, the weaknesses and strengths of children with ASD pair particularly well with the strengths and weaknesses of current AI technology. We take the view that computer-based interaction is a particular kind of *environmental engineering*[26], one in which variability in prosody, word choice, syntactic structure, semantics, and pragmatic context can be systematically controlled and the children's visual and rote memory strengths exploited [6].

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 [1, 19].

2 *Simone Says*

In this section we describe the design of 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 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 lead children through the normal developmental sequence, from Brown's Stage I until early Stage IV [5]. 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 (e.g., kindergarten age). In order to provide practice in language-specific skills and a closer approximation to a true communicative loop, interaction 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 minimal verbal communicative competence, 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. In the future, working with children who are also apraxic may be feasible by replacing the speech recognition component in the current design (SPHINX-II) with an alternate recognizer.

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.

As shown in the figure, the first step is the presentation of a visually engaging but graphically simple stimulus. The core visual and linguistic vocabularies consist of common, everyday objects and actions, both to teach functionally useful language 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 [11]. Although the stimuli are intended to be simple, multiple examples can be 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.

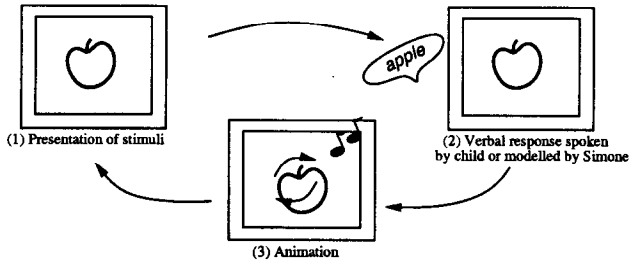

Figure 1: The basic interactive loop

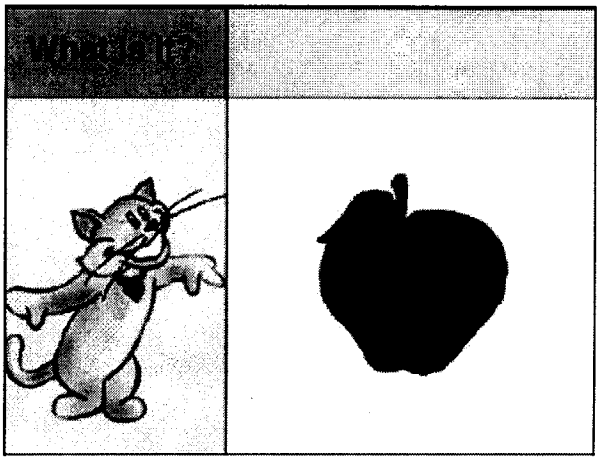

Figure 2: The interface to *Simone Says*

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 more difficult in the normal developmental progression.

Figure 2 shows the interface to *Simone Says* with its four distinct screen areas: Simone's location (lower left) adjacent to the stimuli box (lower right) and text echo boxes for Simone' speech (upper left) and the child's (upper 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 cueing by Simone to indicate the need for a 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 fades in quickly and Simone cues with full body gesture and verbal prompt.

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. Although current technology precludes a purely child-centered teaching style at the level of linguistic phenomena we are trying to support, *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. In all instances, however, only referentially meaningful utterances will produce a response, with Simone modelling an appropriate utterance if the child cannot produce one.

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. Designing the stimuli and animations for *Simone Says* has been done with repeated interaction with therapists and speech/language pathologists. Videotapes of various versions of the stimuli have been shown in small group settings with professionals as well as at the national conference for the Autism Society of America and data has been collected in the form of surveys relating to both general and specific features of the design. The results of this data have fed back into the current version of the system and we anticipate this iterative process to continue (see Section 2.3).

In addition to responding to the feedback we have received from local and national professionals, we can articulate three principles based on more general research in early learning that seem critical to designing the animation sequences:

1. **Make every interaction rewarding.** In other words, playing the game must itself be reinforcing [12, 15]. 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. 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 (a disk that spins in the upper corner of the screen, or a baseball player that advances around bases are examples of reinforcers that are *not* referentially meaningful for the stimuli of an apple but that are, nonetheless, typical of current software design). 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. While part of Simone's role as modeller

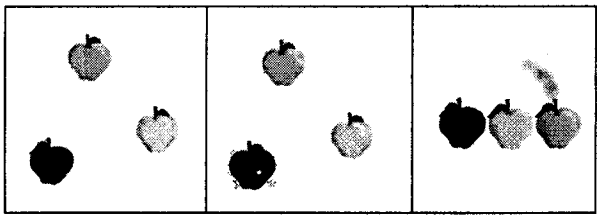

Figure 3: A sequence for introducing the plural form

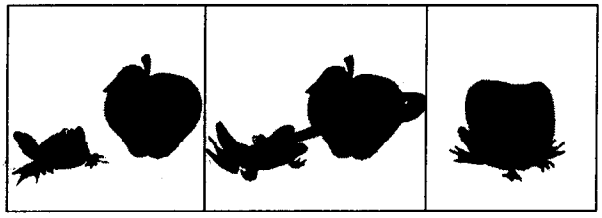

Figure 4: Introducing verbs, and noun/verb pairs

is 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!"), it nevertheless seems useful to treat a few verbs as more generally applicable than they truly are to keep engagement and enthusiasm high.

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 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 (in consideration of space, we omit all but the contents of the stimuli box in this and the remaining figures). 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 as a train (the right frame of 3).

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 modelling of "Yea! Eat the apple!" by Simone.

Movement along the pragmatic dimension requires establishing joint reference with one of the animated characters,

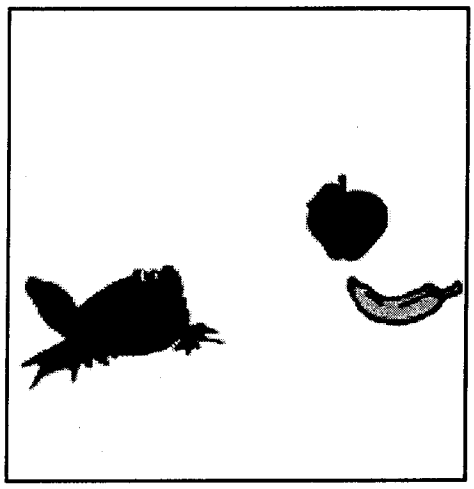

Figure 5: Teaching reference & disambiguation

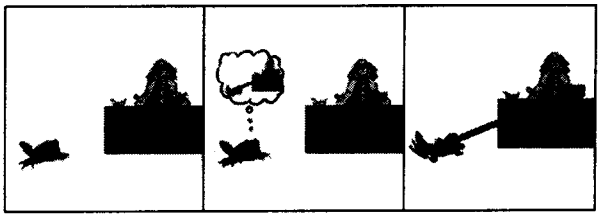

Figure 6: Modelling *theory-of-mind*

an extremely difficult task for children with ASD [16]. Figure 5 shows an initial scene that 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). However, 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?") 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 6 goes a step further by embedding language in a simple social context. Following Baron-Cohen and others [3, 20], 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 secondary 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 short verbal scripts and can be used and reused to target a variety of pragmatic issues, such as point of view ("Take" versus

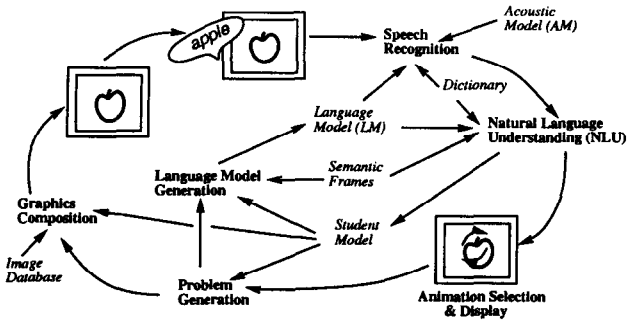

Figure 7: Processing loop and knowledge bases

"Give the apple") and wh-questions, both of which may call for more complex turn-taking behavior and basic 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.

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 [9]). 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, normally these acoustic models would have to be adapted to our target population. Recent work in tutoring reading, however, has resulted in a version of SPHINX-II with an acoustic model adapted to early school-age children [18]. Given that articulation problems and phonemic confusions beyond what is normal for chronological age are not a characteristic of verbal children with ASD, and that the sorts of prosodic differences that are characteristic are filtered out during the initial phases of speech processing, we believe further adaptation of SPHINX's acoustic model will require only a modest additional corpus of speech to be collected (as outlined in Section 2.3).

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.[1] 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 these users.

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 [13]. Starting with a small core grammar and semantic representation for the task domain, CHAMP understands each utterance typed by 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

[1] The LM for a given scene is a function of the stimuli and the student's proximal zone of development, 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. This approach is computationally feasible only because, between Stages I and IV, n always remains very small.

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 success of *Simone Says* relies on combining and expanding existing technology in new ways to meet classical intervention objectives. As such, the project requires evaluation on both technological and pedagogical grounds. As there is a natural dependency we cannot hope to deliver effective intervention if the technology is not up to the task envisioned for it, we consider each in turn.

Evaluating the Technology

Although cheap to reproduce, sophisticated, robust software systems are expensive and time-consuming to build. Thus, identifying and testing underlying assumptions early in the research is prudent. We have identified three important issues regarding the technical 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 stand-alone microphone will not be aversive.

We believe that the best method for testing these assumptions is via an initial "Wizard-of-Oz" experiment in which a mock version of *Simone Says* is used with a human therapist in the loop. This version, already partially constructed, uses the animation database and a simplified problem generator to present the visual stimuli and rewards to the child. The child interacts via speech, as with the real system, but the interpretation of the response (the speech understanding component) is done in real-time by a therapist interacting with the system via a simple touchpad. The touchpad interface limits the therapist to making little more than accept/reject decisions about the the child's utterance in the context of the visual stimulus; since the "wizardry" provided by the therapist involves no special skills, training with the set-up should be straightforward. We anticipate conducting the Wizard-of-Oz experiment in the late spring of 1998 and have agreements of participation from three educational institutions in the western Pennsylvania area.

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. What we expect to learn via this first evaluation is whether the children can work with the technology and if our initial guesses about interface design, stimulus selection and engaging animation are accurate. At the same time, we will be collecting the children's speech during their interactions. A portion of the collected speech samples will be used to adapt the Acoustic Model of the recognizer. Then the examples used in the experiment and the utterances given in response will form development data that can 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.

Evaluating the Intervention

The ultimate goal of *Simone Says* is, of course, to help children with ASD acquire functionally useful language. The sort of short-term evaluations necessary for determining technical feasability are inappropriate for measuring language change. 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* over a one year duration. 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.

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?

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 standard instruments for evaluating productive syntax and pragmatics.

To answer the third question we will use a standard experimental versus control design, with half our subjects receiving intervention with *Simone Says*, and half receiving no software intervention. The dependent variables will include the language test 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* can contribute effectively to that growth.

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 term. 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. 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.

3 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 the independently-developed technologies together into a coherent, engaging, and pedagogically effective real-time environment.

Despite the fact that *Simone Says* is an attempt to integrate pre-existing technologies it would be inappropriate to conclude that this project is simply applied research or product development. We expect the integration of CHAMP and SPHINX to extend our understanding of both adaptive parsing and dynamic, incremental language model generation. We expect to produce interim results in the form of a corpus of child speech data and a database of the developmental sequences of those children with ASD who participate in the longitudinal study. 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 that should be of interest to researchers in cognitive science, language development, and autism.

In addition to these concrete contributions to basic research, we believe that *Simone Says* has the potential to provide a unique platform for collecting data and testing hypotheses that, in turn, can inform our models of human language processing. A computational system makes it practical to systematically examine relationships between language learning and other factors such as rate of repetition, the variability (or constancy) of prosodic, lexical, and syntactic information in the environment, and the importance of non-verbal cues like gaze-following and pointing. It allows us to explore such relationships empirically in well-controlled, reproducible experimental settings. Moreover, an instrumented environment with which children will interact over an extended period of time gives unprecedented access to fine-grained acquisition data. Imagine having ten or fifteen minutes a day of millisecond response times, each day for months, for each of a large number of children who, as a clinically-interesting population, might be distributed sparsely across the country. Such a scenario stands in stark contrast to the more typical longitudinal paradigm of videotaping and transcribing brief samples of language intermittently collected for a small number of children at great expense.

The most compelling reason for this work, however, is what it can mean to children with autism. The lesson from the new therapeutic focus on early intervention is quite clear: acquiring age-appropriate language has a profound effect on behavior, socialization, and the long-term prognosis for an independent adulthood. By providing meaning-based interactive experiences that range linguistically from vocabulary-building to simple social discourse, *Simone Says* may be the first chance for children who need it to learn the efficacy of language from a constantly available, infinitely patient teacher. 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 3-5who enter school with a language disorder.

References

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

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

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

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

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

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

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

[8] M. Heimann, K. E. Nelson, T. Tjus, and C. Gillberg. 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.

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

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

[11] R. Koegel, A. Egel, and G. Dunlop. Learning characteristics of autistic children. In W. Sailor, B. Wilcox, and L. Brown, editors, *Methods of Instruction for Severely Handicapped Students*. Paul H. Brookes, 1980.

[12] R. L. Koegel and J. Johnson. Motivating language use in autistic children. In Geraldine Dawson, editor, *Autism: Nature, Diagnosis, and Treatment*, pages 310–325. Guilford Press, 1989.

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

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

[15] T. W. Malone. Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 4:333–369, 1981.

[16] D. McArthur and L. B. Adamson. Joint attention in preverbal children: Autism and developmental language disorder. *Journal of Autism and Developmental Disorders*, 26(5):481–496, 1996.

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

[18] J. Mostow and M. Eskenazi. A database of children's speech. In *Proceedings of the NSF Interactive Systems Grantees Workshop (ISGW97)*, 1997.

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

[20] S. Ozonoff and J. N. Miller. 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.

[21] B. M. Prizant and A. M. Wetherby. 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.

[22] S. J. Rogers. Brief report: Early intervention in autism. *Journal of Autism and Developmental Disorders*, 26(2):243–246, 1996.

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

[24] S. Steiner and V. Larson. Integrating microcomputers into language intervention. *Topics in Language Disorders*, 11:18–30, 1991.

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

[26] A. M. Wetherby and B. M. Prizant. 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.

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