The Role of Evaluation in Bringing NLP to AAC: A Case to Consider

Kathleen F. McCoy
University of Delaware
mccoy@cis.udel.edu

Dave Hershberger
Prentke Romich Company
dhh@prentrom.com

Abstract

Evaluation of prototype AAC technologies is a very difficult task for several reasons. Among these are the difficulties inherent in evaluating a ``partial'' system -- i.e., one whose focus is on a single aspect of an overall system. For example, for several years, we have been applying natural language processing techniques to the field of AAC in order to develop intelligent communication aids that attempt to provide linguistically ``correct'' output while increasing communication rate. Our focus has been on the processing and system knowledge required in order to expand the user's input. The outcome motivating this project was primarily rate enhancement. While a research prototype was developed at the University of Delaware based on an NLP technique we called COMPANSION (because it takes a COMPressed message and through expANSION, converts it into a well-formed sentence), its practical deployment and outcome evaluation faces several difficulties. These are primarily because the focus of the technique was on processing, but an evaluation requires, and is dependent on, an entire device (i.e., input interface, processing, and output interface). We include an informal experiment which allows a partial analysis of the technique. While such experiments are unable to shed light on possible outcomes of system use, they do validate some assumptions and point out differences among users from different populations.

In continuing our investigation of how COMPANSION might be incorporated into a viable AAC device, a joint project between the University of Delaware and the Prentke Romich Company was undertaken to investigate the possibility of incorporating COMPANSION into a viable communication device for a particular population. The development methodology for this project includes ongoing evaluation of sub-components of the system and tailoring of the system processing to the specific population through a data collection and analysis effort. A portion of the collected data has been set aside for testing purposes.

1. Introduction

There has been a great deal of discussion about outcomes in AAC -- and indeed the measurement of the outcomes of various AAC methodologies with various populations of AAC consumers is a very important research question. One must recognize, however, that a particular instantiation of an AAC prototype device consists of many different components. In addition, there are several different dimensions along which outcomes can be measured. This paper can be viewed as a cautionary note about drawing too strong a conclusion about (either positive or negative) results of evaluating outcomes of a particular AAC methodology. In particular, not only must one identify what kinds of outcomes are of interest, one must also decide which component of the system is responsible for the outcome results.

In a research setting we often conceive of an idea -- generally pertaining to just a portion of an AAC device. However, in order to test the efficacy of this idea, an entire system must be instantiated. One is then left with the question of which component of the system is responsible for a particular evaluation result.

In this paper, we first abstractly describe the components of an AAC system and discuss trade-offs concerning decisions made with respect to these various components. Next we discuss different kinds of measurable outcomes that a particular AAC device may have on a user and indicate how a device may have an unexpected positive outcome along some dimension that was not originally planned for. We then describe a particular research prototype whose intention was to demonstrate the feasibility of a particular natural language processing methodology, and point out the difficulty in evaluating the prototype system. We describe an informal experiment which can be used to validate implementation. However, in order to evaluate outcomes of using the processing methodology, a full system must be developed with a particular population of users in mind. We describe our efforts toward this end with emphasis on our methodology for incremental testing of subpieces of the prototype, and tuning the processing to the particular target population. Hopefully this methodology will indicate various kinds of deficiencies in subpieces of the system which could be remedied before the complete system is evaluated.

2. Computer-Based Augmentative and Alternative Communication

A traditional computer-based AAC system can be viewed as providing the user with a “virtual keyboard” that enables the user to select items to be output to a speech synthesizer or other application. Such a device can be thought of as consisting of four components: (1) a physical input interface providing the method for activating the keyboard (and thus selecting its elements), (2) a language set containing the elements that may be selected; in the language set we must consider what the items are (e.g., letters, words, phrases), and how the items are

organized for selection (e.g., letters in alphabetical order or with most frequently selected letters in front), (3) a processing method which may consist of several levels and is responsible for creating some output depending on the selected items, and (4) an output interface (e.g., a speech synthesizer) which provides feedback to the system user and/or to his/her communication partners. All of these elements must be tailored to an individual depending on his/her physical and cognitive circumstances and the task they are intending to perform.

For example, for people with severe physical limitations, access to the device might be limited to a single switch. A physical interface that might be appropriate in this case involves row-column scanning of the language set that is arranged (perhaps in a hierarchical fashion) as a matrix on the display. The user would make selections by appropriately hitting the switch when a visual cursor crosses the desired items. In row-column scanning the cursor first highlights each row moving down the screen at a rate appropriate for the user. When the cursor comes to the row containing the desired item, the user hits the switch causing the cursor to advance across the selected row, highlighting each item in turn. The user hits the switch again when the highlighting reaches the desired item in order to select it. For users with less severe physical disabilities, a physical interface using a keyboard may be appropriate. The size of the keys on the board and their activation method may need to be tailored to the abilities of the particular user.

Notice that the components of the AAC device are not independent of each other. Consider that the language set low-level processing method may influence the language elements that make up the selectable items. For instance, suppose that the low-level processing in the system consists of word prediction where the system attempts to predict the desired word on the basis of the first several letters. Then the language set had better contain elements that enable the user to select a predicted word or to continue typing out one letter at a time, and these items should be organized in such a way so as to make the selection decision easy. Thus in measuring outcomes, suppose that a word prediction system (a *processing method*) is evaluated and appears to be ineffective in that it does not raise communication rate. Such a failure may not be due to the processing method at all, but it may rather be the result of choices made with respect to the language set and the way that the user may select a predicted word.

One purpose of this paper is to point out places where these types of misleading conclusions can be avoided by separation and individually testing various components of the system. For example, in a case where the theoretical keystroke savings of a word prediction system (when measured appropriately) is high but communication rate is not, one might turn to non-processing aspects of the system in order to find the cause. Presumably such testing might decide which component is at fault when a particular evaluation is not optimal.

Consider as well that selection of components for a computer-based AAC device generally has many trade-offs. Assuming a physical interface of row-column scanning, a language set consisting of letters would give the user the most flexibility, but would cause standard message construction to be very time consuming. On the other hand, a language set consisting of words or phrases might seem more desirable from the standpoint of speed, but then the size of the language set would be much larger causing the user to take longer (on average)

to access an individual member. In addition, if words or phrases are used, typically the words would have to be arranged in some hierarchical fashion, and thus there would be a cognitive/physical/visual load involved in remembering and accessing the individual words and phrases. Each of the choices must depend on the user of the system (i.e., their physical and cognitive abilities) and on the kinds of outcomes desired.

3. Outcomes

In addition to matching the user's physical and cognitive abilities to the system components, one must also consider the kinds of outcomes that may be important for a particular consumer. Like the trade-off in various decisions concerning the system components, there may be trade-offs in outcome possibilities. By the same token, a system designed with one kind of outcome in mind (e.g., rate enhancement) may show an unexpected positive benefit along another outcome dimension (e.g., literacy enhancement).

The outcome of a particular AAC intervention may lie along several different planes. Consider that an outcome to be measured might reflect the **immediate consequences** of the device use. Questions here include whether the device facilitates:

- faster communication?
- better ability to express oneself?
- fewer keystrokes?
- more fluent (natural) conversation?
- more natural interactions?
- longer turns?
- positive perceptions of communicative competence?

One might also consider more **long range consequences** of the device use. Questions here include whether the device has a positive effect on:

- interaction?
- literacy skills?
- turn-taking skills?
- socialization?
- personal opportunities because of improved communication abilities?
- the user's communicative competence?

Finally, one might consider some questions about the **practical usability** or non-communicative aspects of device use. Questions here include does the user:

- have more enjoyment using the device?
- want to use the device?
- participate more in conversations when using the device?
- participate less in destructive behavior after the device is introduced?

Notice that a negative outcome evaluation could be the result of several different things. For example it could be:

- the physical user interface is not appropriate
- the ``editing" facilities provided in the language set are not sufficient for the task
- the language set itself is too complex
- the processing method requires too much cognitive load
- the processing method does not provide a good match with the user's style
- improper training or instruction

Thus, care must be taken when drawing conclusions about outcomes both in terms of the importance of one kind of outcome over another and in terms of determining the true sources of an evaluation result.

4. New Technique: Processing Evaluation

In this paper we focus on some of the results from a research project that has been ongoing for nearly ten years. The beginning of the research culminated in the development of a processing technique known as COMPANSION. Because the COMPANSION project was focused on processing and did not consider the other components of an AAC system, a full evaluation of the technique was not possible. Instead, the technique was informally evaluated through a simulation experiment (described below). Through this experiment several challenges with bringing the technique to a full AAC system were uncovered. Subsequent to this, the team at the University of Delaware joined up with a team from the Prentke Romich Company. The expertise of these two teams lies in different parts of the total AAC system. We describe our current research effort with emphasis on our methodology for testing system components in isolation (with respect to a particular user population). In this way we hope to tailor the processing methodology, the language set, and the physical interface to facilitate positive outcomes. The testing methodology (mixed in with the development effort) has the potential for saving a great deal of time.

A large research effort at the University of Delaware resulted in the development of a technique that could expand telegraphic sentences into full sentences (McCoy et al., 1989), (Demasco and McCoy, 1992), (McCoy et al., 1994). The technique, termed COMPANSION (because it takes a COMPressed message and through expANSION, converts it into a well-formed sentence), was

an effort that concentrated on the Processing phase of an AAC device. Other researchers developed systems with similar goals such as (Hunnicutt, 1986) and (Reich and Shein, 1990). The processing phase itself is rather complicated and requires a great deal of information to be associated with the lexical items (words) that can be selected by the user.

The Processing model underlying COMPANSION was implemented in a prototype system, but little effort was placed on components of the system beyond the processing component. An assumption of the system was that the input interface to the system would be word-based. That is, each word of input would take a (basically) constant amount of time (regardless of how many characters were in the word). We call this constant amount of time a *keystroke*. Word endings (e.g., +s plural or +ed past tense) would require an additional keystroke to select. No more specific assumptions about the physical interface or language set were made.

Thus the focus of the research was on a ``black box'' which took the words input by the user and expanded them into full sentences to be output via an output interface (e.g., print or a speech synthesizer). COMPANSION potentially increases the communication rate by requiring fewer words to be selected (since it requires just the content words of the desired utterance to be input) and by eliminating the need for selecting morphological endings.

As an example of the kind of processing that could be done, consider the following example handled by the prototype research system:

Input: `think red hammer break John`
Output: `I think that the red hammer was broken by John.`

4.1. Evaluating the Technique

A problem facing the evaluation of the COMPANSION technique was that an AAC technique such as COMPANSION cannot really be tested in isolation: it is a high-level processing technique and all other components of the system must be designed and implemented in order to actually test a processing technique. Moreover, once an entire system is implemented, one runs the risk of negative evaluations being a result of a mismatch between the technique (or user) and the choices made for the other system components.

Upon further inspection, we decided to run some experiments that we felt might help evaluate the coverage of the technique itself. In the COMPANSION technique a primary emphasis was the inclusion of a sophisticated semantic knowledge base and numerically-based heuristics for reasoning about relative word roles. For example, the system might take a set of input such as: "`<apple> <pear> <eat> <john>`" and generate "An apple and a pear were eaten by John". Note that in order to generate such a sentence the machine had to recognize that `<apple>` and `<pear>` were the things being eaten (recognizing a conjoined theme), and that John was doing the eating. In addition, appropriate determiners (e.g., "a") were added (but not to proper nouns such as John), and the appropriate passive construction was used (requiring the past tense form of

"be" and a past participle ending on the main verb) in order to maintain the input order selected by the user.

One kind of evaluation one can do with such a technique is to evaluate the inferencing methods of the technique. To accomplish this, we need a methodology for deciding the specific functionality (i.e., the input/output requirements an optimum COMPANSION system should exhibit). Ideally, we would like our system to act like a familiar human partner does. Thus, our initial evaluation attempted to uncover interaction patterns that occur between an AAC user and a familiar listener. Our analysis emphasized the types of linguistic transformations performed in translating word sequences to sentences.

4.2. Method

Pilot data was collected by transcribing videos originally recorded in conjunction with van Balkom. Adolescent students with cerebral palsy described pictures in a children's book to their primary speech therapists, using their own manual symbol charts. Four such adolescent-therapist dyads were videotaped and analyzed.

Each student was instructed to describe the pictures as if telling a story to younger children. The therapist was instructed to repeat each word as it was selected by the student, paraphrase the sentence when it was completed, and then ask the student for confirmation that the paraphrased interpretation was correct. A single camera was used to videotape both the student and the therapist. Students took between 11 minutes and one hour to retell their stories.

4.3. Results

4.3.1. Some Interactions Consistent with the COMPANSION Approach

Standard COMPANSION: Some interactions with the therapist followed the "standard" operation of the COMPANSION system[1].

S: <girl> <make> <in> <pan> <egg> <breakfast>
T: Girl *will* make *the* egg*s* in *the* pan *for* breakfast.

Here the therapist has added tense, and determiners. In addition, the plural form of "egg" was chosen. Though not indicated by the student, the plural form may have been chosen using default knowledge (that people generally eat multiple eggs for breakfast) or it may have been the result of extra-linguistic information (e.g., the picture being described at the time). Notice that the preposition "for" was also included in the expanded message. This addition required reasoning about the semantics of the input sequence. For example,

[1] In this and subsequent examples "S" stands for the student input and "T" the therapist. Words/letters added by the therapist are in italics. Words of particular interest are in bold.

breakfast was the “reason” for making the eggs and should be introduced with a for preposition.

Word Order Changes: An assumption of the COMPANSION system has been that the words will be given to the system in the same order that they should be output in a sentence. However, some of our analysis reveals that the therapist sometimes did not follow the word order initially given by the student. The above example falls into this category: the eggs and the pan have been switched in the therapist's output. Consider the following example as well:

S: <boy> <table> <dusting> <grandmom> <floor> <sweep>
T: Boy *is* **dusting** *the* **table** *and the* grandmom *is* **sweep*ing*** *the* **floor.**

Notice that in this instance the student is not following a standard subject-verb-object ordering of the words. The therapist changes the order to follow standard English word order (it is not obvious how to form an English sentence while keeping the word order given by the student).

Agent Inference: The COMPANSION system expects that a user might omit an agent when referring to him/herself. An agent might also be omitted if it was obvious from context. This behavior was also found in our analysis. Because the story was about a boy and a girl, students sometimes did not specify an agent, yet it was inferred by the therapist:

S: <wash> <clothes>
T: ***They*** *are* wash*ing* clothes.

Verb Inference: Another assumption of the COMPANSION system is that the main verb may be left out in some situations (particularly when the main verb is either have or be). We have argued previously that a system must have the ability to reason about which verb is most appropriate in the given situation. Our default rule (i.e., if there is an animate agent and an inanimate object, then the verb “have” should be inferred) is consistent with examples found in the transcripts. Consider the following where both the agent (“they”) and the verb (“have”) have been inferred.

S: <toys>
T: ***They have*** toys.

Conjunctions and Possessives: Students sometimes left out conjunctions in the pilot study. This occurred both at the sentence level and in both subject and object positions in the sentence:

S: <boy> <girl> <made> <bed> <up> <in> <morning>
T: *The* boy ***and*** *the* girl made up *the* bed in *the* morning.

The inference of when a conjunction is necessary is complicated by the need to correctly indicate possessive information. The following example contains an inferred possessive.

S: <boy> <girl> <girl> <clothes> <my> <mother>
T: ***They're*** *giving* ***their*** clothes *to* ***their*** mother.

This example is interesting in that it points out several of the difficulties inherent in inferring when a possessive is needed. Note above there was both a conjunction (<boy> and <girl> combined to “they”) and two possessives. A possible possessive rule might require that if you want a possessive followed by a noun, just put the two items next to each other (e.g., <girl> <clothes> for “the girl's clothes”). Note here <girl> <clothes> was translated as “their clothes” as if <girl> was now “standing for” the combined agent. However, this strategy was not followed for the second possessive (the strategy would have resulted in <girl> <mother> being used). Rather the student chose the first person possessive pronoun, “my”, to indicate the recipient in the message.

It is not clear in the data how much of the therapist's interpretation was influenced by the picture book itself. Nonetheless, this example raises important questions about how to determine when a possessive form is desired.

4.3.2. Some Interactions Beyond the Scope of Current Technology

Dropped Word (in interpretation): In some instances the therapist did not include words given by the student in the interpretation even though they often contributed to the intended meaning. Consider:

S: **<two>** **<table>** <in> <table> <dining> <room>
T: *There were things on the* table in *the* dining room.

Notice that table occurs twice in the student's input, but only once in the interpretation. In some sense, the student's input is “linguistically” sound. He is saying two things about a table (a) there are two things on it, and (b) the table is in the dining room. If these two assertions were stated as two separate sentences, then “table” would occur twice. However, as a single sentence there is a way to combine the thoughts without repeating “table”. Compare this example with the possessive case above for an illustration of the difficulty in distinguishing this case from that of a possessive.

Other times the dropped word did not contribute to the meaning:

S: <girl> <look> <at> **<to>** <boy>
T: *The* girl*'s* look*ing* at *the* boy.

Replacing a Word (not included in interpretation): In some instances the therapist ignored words selected by the student, even though there was no obvious indication from the student to ignore the word.

S: <girl> **<help>** <clothes> <up>
T: Girl clothes up. She*'s* ***hanging*** *the* clothes up.

Note that in the above example <help> does not occur in the output. The example also shows a case where a new verb has been inferred (probably from the extra-linguistic context).

More Complicated Verb Inference (Adding or Replacing a Word): In some instances the therapist inferred a verb which was not actually included in the input:

S: <boy> <girl> <up> <table> <for> <lunch>

T: OK. *They're* ***setting*** up *the* table for lunch.

4.4. Discussion

A study such as that described here has some advantages, but also raises some questions. Presumably it gives us insight into the limits of a proposed technique and indicates what we should strive for in an implementation.

What this study does not tell us is whether or not the technique is effective or whether an entire, usable system that uses the technique can be built. In addition, it is likely that different populations of users will use different linguistic structures in their expression. Even here we must take care in the conclusions drawn. For instance:

- We do not know what prior knowledge led to the therapist's interpretation. For instance, since both participants saw the pictures, did prior knowledge of anticipation play a major factor in determining the communicative intent? For example, inferring 'they' as the subject instead of 'I', 'he', 'she' etc. would be much easier while looking at a picture of two people. This same sort of thing is likely to occur when a familiar listener and an AAC user have shared knowledge concerning a situation.
- We do not know what kind of interaction between a machine and a user might be appropriate. In the experiments the students were performing a task given to them by a therapist. If the therapist misinterpreted their intent, would they be willing to try to 'correct' the therapist or would they be content that an acceptable answer had been provided? This may be different when a non-speaking person initiates communication in order to convey information to a person.
- We do not know whether the telegraphic speech is a result of “intentional”. omissions (due to a conscious decision on the AAC user's part or due to their language abilities) or whether selections to create fully grammatical forms were not available on the system. In other words, were articles, conjunctions, etc. omitted to increase rate, omitted because the individual didn't know how to use them properly, or omitted because those words did not exist on their communication board and therefore could not be generated? We should draw different conclusions (and perhaps different kinds of interventions) depending on which of these was the case.

Thus this experiment gives us some insight, but the development of a specific system for a specific task is ultimately necessary. Since during this development many decisions may affect the ultimate usability of the system, one must (1) choose a specific population, and (2) tailor all system components to individuals to ensure system usability.

5. Joining Forces: Expertise from Several Places

The University of Delaware and the Prentke Romich Company joined forces in order to develop a prototype system that used the COMPANSION technique and contained all of the other system components. In order to be able to tailor all of the system components, we focused on a single target population and attempted to build system components that would be appropriate for this population. At each stage, individual testing of components has been an important step.

5.1. Target Population

In considering a target population we looked for a group of users who would likely produce telegraphic input, would benefit from the expansion of that input into full sentences, and that could be counted on to use a fairly limited number of words and linguistic structures. This was crucial because the COMPANSION technique requires a lot of information on each word and must understand all sentence structures. We chose to consider a young population of users who have cognitive impairments that affect their expressive language ability. Whether a child with cognitive impairments is verbal or nonverbal, their expressive language difficulties may include the following (Kumin, 1994), (Roth and Casset-James, 1989): (1) short telegraphic utterances; (2) sentences consisting of concrete vocabulary (particularly nouns); (3) morphological and syntactical difficulties such as inappropriate use of verb tenses, plurals, and pronouns; (4) word additions, omissions, or substitutions; and (5) incorrect word order. While such children may have the ability to functionally communicate their needs and wants, intervention to assist them in their language production should be beneficial both from a social and an educational perspective.

In developing a device geared toward this population, several issues must be dealt with. Here we focus on three to emphasize our processing methodology:

lexical access -- what is an appropriate method for providing such a user with access to the lexical items that they wish to communicate? This includes language set and physical input interface.

verification of user input/output assumptions -- what kind of input will this population produce and what expansions are reasonable?

user interface issues -- what kind of interface is necessary for a user with cognitive impairments to be able to access the system? Crucial here is the user's ability to sift through the expansions provided by the system and select the one they desire for output.

5.2. Lexical Access: Communic-Ease Map™

PRC has a great deal of expertise in the area of physical input interfaces appropriate for a wide variety of users. In addition, they have provided effective

language sets coupled with low-level processing to provide users with a mechanism for outputting desired messages. In fact PRC has expertise in providing lexical access to the population under study. The speech output communication aids that PRC designs for commercial use incorporate an encoding technique called semantic compaction, commercially known as Minspeak® (a contraction of the phrase ``minimum effort speech") (Baker, 1982), (Baker, 1984). The purpose behind Minspeak® is to reduce the cognitive demand as well as the number of physical activations required to generate effective flexible communication. It uses a language set (i.e., a set of selectable items) consisting of a relatively small set of icons that are rich in meaning and associations. These icons can be combined to represent a vocabulary item such as a word, phrase, or sentence, so that only two or three activations are needed to retrieve an item. This small set of icons thus allows access to a large vocabulary which is stored in the device. Since they are rich in meaning, icons designed for Minspeak® can be combined in a large number of distinct sequences to represent a core lexicon easily.

The Minspeak® language set and processing was first utilized with PRC's Touch Talker™ and Light Talker™ communication aids (which united different physical interfaces with the icon encoding). With these Minspeak® systems, if icons on the overlay remain in fixed positions, once learned, they allow the individual using the system to find them quickly and automatically. This automatic processing was facilitated by the design of prestored vocabulary programs known as Minspeak® Application Programs (MAPs™). In these programs a large vocabulary is prestored in a well-organized fashion using a logical, paradigmatic structure that greatly facilitates learning and effective communication.

One of these MAPs™, Communic-Ease™, contains basic vocabulary appropriate for a user chronologically 10 or more years of age with a language age of 5-6 years. Communic-Ease™ has proven to be an effective interface for users in our target population providing access to approximately 580 single words divided into 38 general categories. Most of these words are coded as 2-icon sequences. The first icon in the sequence (the category icon) establishes the word category. For example, the <SKULL> icon indicates a body part word, the <MASKS> icon indicates a feeling word, and the <APPLE> icon indicates a food word. The second icon denotes the specific word. For example, <MASK> followed by <SUN> produces the word “happy”; <APPLE> followed by <APPLE> produces the word “eat”.

In addition to the words which are accessed via the icon sequences, Communic-Ease™ contains some morphology and allows the addition of endings to regular tense verbs and regular noun plurals. However to accomplish this, additional keystrokes are required. It is also possible to spell words that are not included in the core vocabulary.

The Communic-Ease™ MAP has proven to be an effective means of communication for individuals in our target population. Thus this MAP™ implemented on PRC hardware has provided a physical input interface, the portion of the language set necessary for selecting vocabulary items, and a low-

level processing technique which allows users within the target population to functionally communicate. However, users tend to produce telegraphic messages consisting of key word sequences. Thus they would likely benefit from the addition of COMPANSION-like processing.

5.3. Design Methodology: User Centered Design

Notice that the Communic-Ease™ MAP and the PRC input/output interfaces are proven useful for the population under study. We take these components, but must ensure that the remaining components are appropriate for the user population. Some issues we must handle include:

- What is the range of input structures the system must handle?
- What are the appropriate expansions of that input (and how can the machine be programmed to output a set of appropriate expansions)?
- What additions must be made to the language set to provide appropriate selection and editing facilities?
- How should the language set be organized?
- What kind of interface will allow effective use of the system?

Our methodology in this collaborative effort is to design a system that is geared toward the specific user population. Thus, we have set out to validate our assumptions about the user input and output requirements and to tune the user interface (i.e., editing, language set organization, and physical interface) to the population. Our system input/output functionality has been determined by a collection of transcripts from Communic-Ease™ users. We have collected both raw keystroke data (so that we can establish the range of input we expect from the population) and keystroke data from videotaped sessions where interpretations of the keystroke data are provided by a communication partner. This data allows us to ensure the output from the system is in fact appropriate.

Collection of such data has allowed us to:

- validate expected sentence structures
- validate the expectation of limited vocabulary
- validate input assumptions

In addition, we plan to validate our interface requirements on the basis of iterative user testing. The interface will be developed so that it can be customized to the specific needs of particular users.

5.4. User Interface Issues: User Centered Design

5.4.1. Envisioned System

The envisioned system combines the PRC Liberator™ system (which provides both a physical interface and low level processing) that runs a modified

Communic-Ease™ MAP(which provides a standard vocabulary and its access method) and an intelligent parser/generator (which provides the COMPANSION-like processing). The input from the user will be through the Liberator™ keyboard (most of whose keys contain the icons which are transformed into words via the Communic-Ease™ MAP). The user will receive feedback through an Interface Display. Part of the Interface Display looks much like a standard Liberator™ display (e.g., showing selected icons and words). An additional area of the Interface Display will show the transformed sentences which the user may select to be "spoken" by the system. The user may ask the system to speak the sentences through a private audio channel in cases where s/he is unable to read the display.

The Liberator™ Overlay/Keyboard accepts user input via a variety of methods (e.g., direct selection), and can also limit user choices via Icon Prediction. With Icon Prediction only icons that are part of a valid sequence are selectable. The user selects icon sequences that are transduced into words or commands according to the Communic-Ease™ MAP. In normal operation, icon labels and the transduced words are sent to the Interface Display to give the user feedback (words may also be spoken incrementally).

In the proposed system, these components are supplemented with an intelligent parser/generator (IPG) that is currently under development. IPG is responsible for generating well-formed sentences from the user's selected words and is a simplified version of COMPANSION. IPG also provides further constraints on the Icon Prediction process. For example, if the user selected "I have red". the system might only allow icon sequences for words that can be described by a color (e.g., shoe, face).

IPG encodes a set of transformations for expanding sentences input by the user. These transformations have been motivated by our study of transcripts collected from current Communic-Ease™ users. Using these transcripts the processing of the system can be tuned to handle the kinds of constructions common to this particular population.

5.4.2. Interface Issues: Isolation and Testing

Beyond the basic operation described above, there are a number of interface issues that need to be resolved before a completed system is developed. These issues are being explored in early interface prototypes with iterative user testing. For example, one question with the interface concerns the method with which users select the desired expansion when the system comes up with several possibilities. The particular user population poses several challenges; they most likely cannot read, and may not be able to remember what they desire if several possibilities are presented to them.

Our methodology here includes building a prototype interface (using our intuitions from knowledge of the target population) and then adjusting the interface through user testing and further interface development. This interface testing need not be done in the context of this particular system. Rather a "simpler" task will be given (e.g., a game) to test the feasibility of the interface

components. In this way we hope to isolate interface components from the cognitive demands of the entire system.

For instance, one important aspect of the system operation is selecting the desired expansion from a list of possibilities calculated by the system. We have implemented an interface which provides for selecting from a list which is presented to the user both visually and auditorially. The system "scans" through the list one at a time highlighting the list item and speaking the item through a private auditory channel. The user may select the desired item at any time during the scan. The interface is designed flexibly. The number of items to select from, the speed of the scan, and the method of selection may all be customized by changing some system parameters.

Rather than testing this interface in the context of the whole system (which may be confusing for a user), the interface testing is planned in a simple "game" situation. In the game the user will have to select the correct item from a list. In this way the selection and presentation parts of the interface may be tuned to the user in isolation from the rest of the system. Other aspects of the interface (e.g., the editing functionality, the presentation layout) will be developed in a similar manner.

5.5. Development Methodology: Verifying User Input/Output Assumptions

The prototype system combines the PRC's Liberator™ platform and Communic-Ease™ Map with the current-generation intelligent parser/generator. In the implementation the Liberator™ will function primarily as the user's keyboard and a tablet-based portable computer will contain the parser/ generator and function as the Interface Display. The two systems will be connected via an RS-232 or IR link. This strategy allows for rapid initial prototype development.

Our project methodology is to develop and test the robustness and usability of the system in phases. The parser has been developed in C++ and is being refined and tested as other parts of the project progress. A core grammar has been created and is being revised and enhanced to handle a larger variety of structures. Current lexicon efforts involve expanding the number of entries beyond the basic Communic-Ease™ vocabulary and adding the necessary semantic knowledge. The first version of the Windows-based user interface has recently been completed and is now being evaluated internally.

Several evaluations of the completed prototype system are planned. For instance, a theoretical evaluation of the grammar coverage is ongoing. As has been stated, we have collected key selections from current users of the Communic-Ease™ MAP. In some situations, we also have an interpretation of those keystrokes provided by the communication partner in a videotaped session. These video sessions have been transcribed and aligned with the keystroke data. While some of this data is being used to develop the grammar, we have set aside a portion of it to be used for testing purposes. This test data will allow us to test the system's grammar in several ways. First, the robustness of the grammar can

be tested by determining the number of completed input utterances found in the collected data that can be handled by the grammar. Second, the appropriateness of the grammar can be tested by determining how often the grammar's output matches the interpretation provided by the communication partner in the video sessions. Because we have much more keystroke data than transcribed video data, we also plan a test of grammar appropriateness by comparing the output of the grammar with that generated by a human faced with the same sequence of words.

In addition to the theoretical grammar testing described above, we also plan an informal evaluation of the usability of the system. We plan to iteratively refine the interface by doing usability studies of our prototype with current users of the Communic-Ease™ Map. One aspect these studies may shed light on is whether or not users in the population under study can in fact select their desired sentence when a list of possibilities is presented to them.

6. Conclusions

Evaluation of a new AAC methodology is a very difficult task. A complete AAC system consists of multiple components -- all of which must be tuned to particular users. In evaluating outcomes, care must be taken to appropriately determine which component must be updated. Any one of the system's components may be responsible for a negative outcome and one must take care that conclusions drawn are not too broad. In addition, a negative outcome may be the result of poor training or not enough practice on the system. Not only this, but the outcomes themselves contain a great deal of variety. For instance, a system initially conceived as a rate enhancement technique may end up having a positive effect on literacy skills.

Here we focused on the development of one particular project and have discussed the notion of separate evaluation of subcomponents and on-going evaluation in conjunction with the development.

Past efforts have allowed us to take some components which have already proven useful for this population. For instance we know that the target population is already accustomed to the access technique, the vocabulary, and the language encoding system of the Liberator™. We have described ways of testing subcomponents (e.g., the processing sophistication of IPG). However, new issues come up in integration which require further attention. These include questions such as: "How cognitively disorienting is the additional information provided by the system?", and "If additional information is provided, how should it be presented?". Questions such as these provide avenues for further work.

7. Acknowledgments

This work has been supported by a Small Business Research Program Phase I Grant from the Department of Health and Human Services Public Health

Service, and a Rehabilitation Engineering Research Center Grant from the National Institute on Disability and Rehabilitation Research of the U.S. Department of Education (H133E30010). Additional support has been provided by the Nemours Foundation.

The authors would like to thank Arlene Badman, Patrick Demasco, Clifford Kushler, and Christopher Pennington for their collaboration on the project. In addition we thank John Gray for his discussions and implementation of many of the C++ aspects of the system, and Marjeta Cedilnik for her work on the grammar (and transformation rules).

8. References

B. Baker. Minspeak. *Byte*, page 186, September 1982.

B. Baker. Semantic compaction for sub-sentence vocabulary units compared to other encoding and prediction systems. *In Proceedings of the 10th Conference on Rehabilitation Technology*, pages 118--120, San Jose CA, 1984. RESNA.

Patrick W. Demasco and Kathleen F. McCoy. Generating text from compressed input: An intelligent interface for people with severe motor impairments. *Communications of the ACM*, 35(5):68--78, May 1992.

S. Hunnicutt. Bliss symbol-to-speech conversion: 'Blisstalk'. *Journal of the American Voice I/O Society*, 3, 1986.

L. Kumin. *Communication Skills in Children with Down Syndrome: A Guide for Parents*. Woodbine House, Rockville, MD, 1994.

K. McCoy, P. Demasco, Y. Gong, C. Pennington, and C. Rowe. Toward a communication device which generates sentences. In *Proceedings of the 12th Annual RESNA Conference*, New Orleans, Louisiana, June 1989. RESNA.

Kathleen F. McCoy, Patrick W. Demasco, Mark A. Jones, Christopher A. Pennington, Peter B. Vanderheyden, and Wendy M. Zickus. A communication tool for people with disabilities: Lexical semantics for filling in the pieces. In *Proceedings of the First Annual ACM Conference on Assistive Technologies (ASSETS94)*, pages 107--114, Marina del Ray, CA:, 1994.

F. P. Roth and E. Casset-James. The language assessment process: Clinical implications for individuals with severe speech impairments. *Augmentative and Alternative Communication*, 5:165--172, 1989.

Peter Reich and Shein F. VOICI: a voice output intelligent communication system. Presented at ISAAC-90, volume 6. ISAAC, Abstract in AAC -- Augmentative and Alternative Communication, 1990.