Int. J. Man–Machine Studies (1991) **34**, 527–547

How to get people to say and type what computers can understand

ELIZABETH ZOLTAN-FORD

Department of Psychology, Towson State University, Towson, Maryland 21204, USA

(Received 9 August 1988 and accepted in revised form 1 July 1989)

This study tested whether people can be shaped to use the vocabulary and phrase structure of a program's output in creating their own inputs. Occasional computer-users interacted with four versions of an inventory program ostensibly capable of understanding natural-language inputs. The four versions differed in the vocabulary and the phrase length presented on the subjects' computer screen. Within each version, the program's outputs were worded consistently and presented repetitively in the hope that subjects would use the outputs as a model for their inputs. Although not told so in advance, one-half of the subjects were restricted to input phrases identical to those used by their respective program (shaping condition), the other half were not (modeling condition). Additionally, one-half of the subjects communicated with the program by speaking, the other half by typing. The analysis of the verbal dependent variables revealed four noteworthy findings. First, users will model the length of a program's output. Second, it is easier for people to model and to be shaped to terse, as opposed to conversational, output phrases. Third, shaping users' inputs through error messages is more successful in limiting the variability in their language than is relying on them to model the program's outputs. Fourth, mode of communication and output vocabulary do not affect the degree to which modeling or shaping occur in person–computer interactions. Comparisons of pre- and post-experimental attitudes show that both restricted and unrestricted subjects felt significantly more positive toward computers *after* their interactions with the natural-language system. Other performance and attitude differences as well as implications for the development of natural-language processors are discussed.

Introduction

Petrick (1976) summarized the status of natural-language interfaces. He stated that critics of such interfaces argue that people's natural-language inputs are too ambiguous and loosely structured for computers to understand. After careful examinations of such inputs, however, Petrick and others have found that natural-language inputs are not ambiguous, but rather are clear and relevant to the users' task (Malhotra, 1975; Harris, 1977; Zoltan, Weeks & Ford, 1982). What is problematic about natural language is the freedom it allows its users. People use a variety of words and syntactic structures to request the same response from a computer (Petrick, 1976; Ford, 1981; Zoltan *et al.*, 1982). The low recognition rates of the majority of natural-language processors are a result of their inability to handle this variety.

There are two ways to maximize the probability that a natural-language processor will understand each user input: (1) program the computer to understand the many ways people can structure their inputs; or (2) curtail the variability in people's

0020-7373/91/040527 + 21$03.00/0

inputs. In the first instance, the computer must accomodate the user. To do so, the developer of a natural-language processor must anticipate not only which words people will choose, but also how these words will be strung together. Unfortunately, this approach is not particularly useful. There will always be one more way for a user to phrase a request that the designer did not anticipate.

A second approach is to have the users adjust themselves to the computer's limited understanding. This approach can be handled in two ways: overtly or covertly. Overtly, people can be given a limited set of acceptable words or phrases to use when they communicate with the computer. Opponents of natural-language processing contend, however, that placing such a restriction on people's natural language is too cumbersome for the user (Petrick, 1976). This argument is based on the belief that synonomous, natural ways of stating a request will interfere with the user's recall of the restricted set of allowable inputs (Black & Moran, 1982). In doing so, restricted natural language becomes difficult to learn and frustrating to use (Kelly & Chapanis, 1977).

Another alternative is to restrict people's natural-language covertly. According to Becker (1975), people have a phrasal lexicon consisting of six major categories of lexical phrases. In creating their natural-language communications, people refer to this lexicon and "stitch together swatches of text that [they] have heard before" (p. 38). The goal in person–computer interactions is to capitalize on this *stitching* process so the combined *swatches* remain consistent both within and among a system's many users.

The stitching process might be accomplished by limiting the language used by the computer program. Natural-language programs could be designed to use an output consisting of a restricted group of words and sentence structures. Given that the program's phrases are worded consistently and can be transposed to allow users to perform the actions they need, users may stitch together swatches of the program's output. More prosaically, the vocabulary terms and phrase structures presented by the computer may serve as a model for the users' inputs.

The computer-as-model effect does appear to influence people's interactions with computers. For example, consider the unrestricted natural-language of users of Ford's (1981) CHECKBOOK program. Seventy-eight percent of the users' messages were declarative statements such as "I want to enter some cancelled checks" (Zoltan *et al.*, 1982). This finding is surprising because such inputs do not occur frequently in person–computer interactions (Thompson, 1980). According to Thompson, user commands in natural-language interactions are more typically structured as explicit imperatives, such as "enter some cancelled checks".

Perhaps CHECKBOOK's style of communication was modeled by its users. CHECKBOOK continually displayed lengthy, conversational statements and questions (e.g. "The first thing I'm going to do is tell you some of the things that the program can do for you" and "What else do you want me to do?"). Sarason (1957) stated that in a novel situation people will actively seek cues from their surroundings. CHECKBOOK's users seem to have done so.

The research literature on person to person communications further supports the occurrence of modeling effects in verbal exchanges. According to Danzinger (1976), people frequently and unknowingly model the style and content of their partner's speech. Matarazzo, Weitman, Saslow and Wiens (1963), for example, revealed the

occurrence of what they term the "speech duration effect". Along with their colleagues, they have shown repeatedly that the length of an individual's response when interviewed is positively correlated with the length of the interviewer's statement or question (Matarazzo, 1962; Matarazzo, Saslow, Matarazzo & Phillips, 1958; Matarazzo, Hess & Saslow, 1962). Even President Kennedy experienced this effect: analyses of 61 news conferences given by President Kennedy revealed a positive correlation between the length of the reporters' questions and the length of the President's responses (Ray & Webb, 1966). Dyads also model one another's specificity frequencies (using "a" instead of "the"), interpersonal orientations (using "I" instead of "you"), utterance lengths (Jaffe, 1964), and pause lengths (Cassotta, Feldstein & Jaffe, 1968).

Assuming that people will model the characteristics of the computer's output, one way to curtail the variability of people's natural-language inputs is to limit the language used by the computer itself. But if the correspondence between the computer's and the users' language is not perfect, the developer of a natural-language processor may not want to rely solely on the occurrence of modeling. Another method of curtailing the variability of user's language then would be needed.

The area of verbal shaping lends itself to this end. In addition to providing the user with a consistently-phrased output, the computer program could shape users to limit the vocabulary and the grammatical structures of their inputs. Verbal conditioning (shaping) experiments involve attempts to modify the verbal behavior of subjects by selectively reinforcing a particular spoken or written response or response class. Researchers have used such procedures to encourage people to utter plural nouns (Greenspoon, 1955; Mandler & Kaplan, 1956), to use "I" and "we" instead of other pronouns (Taffel, 1955), to use specific types of verbs (Sarason, 1957), to limit the content of their speech either to animals (Ball, 1952) or to mothers (Mock, 1962), to increase the rate of their speech (McNair, 1957), to conform to the attitude of another (Hildum & Brown, 1956), and to show self-acceptance (Nuthmann, 1957).

To understand how this process could work in person–computer interactions, consider the following scenario. A user sits before a computer terminal and presses a single button to start the program. The program greets the user and begins to explain what it can do. The computer's explanation has three noteworthy features. First, it uses the same vocabulary each time it refers to a given action or descriptor. For example, the computer consistently uses the word *delete,* rather than a variety of synonyms such as *erase, expunge, drop, cancel, reduce* or *rid,* to refer to the process of removing a file from the program's memory. Second, phrases and words that are not contingent on any specific action remain consistent across the various descriptions. And third, the computer's explanations are all worded in the second-person future tense.

After the computer completes its description, the user enters his or her first request. Unknown to the user, the computer only understands a limited number of vocabularly words and grammatical structures. Inputs that the computer understands are transpositions of its own outputs from the second-person, future tense to the first-person, present tense. When the user enters requests that are of this form, the computer provides the desired response. The computer's response thus serves to

reinforce the user to continue to create the same types of inputs. Should the user enter a request that does not meet this requirement, the computer cannot interpret it and the user receives an error message. Because the computer cannot understand and respond to these entries, the user's tendency to generate messages like them should decrease. After continued interactions with the computer, the user's language is shaped to become like that of the computer's output.

Statement of problem

The purpose of this research was to test the feasibility of reducing the variability in people's natural language through modeling and shaping. Forty-eight occasional computer users interacted with four versions of an inventory control program ostensibly capable of understanding natural-language inputs. Although identical in the actions they could perform, the versions differed in the familiarity of the vocabulary and the length of the output phrases presented on the subjects' computer screen. Before the users took control of the interactive dialogue, each inventory program described its purpose and capabilities repetitively using its set of vocabularly words and phrase structures. When the program's introduction was concluded, the users interacted with the program to create, manipulate and retrieve inventory file information. One-half of the subjects were unknowingly restricted to inputs that mirrored the linguistic characteristics used by their respective inventory program, the other half were not. Additionally, one-half of the subjects communicated with the program by speaking, the other half by typing.

The linguistic characteristics of the users' inputs and the users' responses to 15 semantic differential attitude scales were examined to answer the following questions:

(1) Will users of such programs naturally accept their program's communication style as their own, in this way modeling the program's language?
(2) If they do not and the computer responds that it cannot understand their messages, then will they choose to rephrase their messages with the words and grammatical structures presented to them by the program? That is, can the computer verbally shape users to restrict their natural language?
(3) Given the rapid advances in voice input technology, are modeling and shaping effects as likely to occur when users speak, instead of type, their messages?
(4) And, how are users' attitudes toward computers affected by their interactions with such programs?

Method

SUBJECTS

Subjects for the study were 48 adults who worked in the Baltimore area business community. Of these 48, 26 were male and 22 were female. The subjects' occupations varied widely and included elementary school teachers, geologists, business-forms and insurance sales people, secretaries, and professional musicians. Although many of the subjects had prior exposure to commercially available software packages and video games, none of the subjects considered themselves experienced in computer programming or computer use.

The experimental design included three replications within each of 16 experimental treatments. The resulting 48 sessions were randomly ordered. Subjects were assigned to experimental conditions on the basis of this randomly ordered session sequence. All subjects were paid for their time.

APPARATUS

Subjects communicated with an IBM 3081 model D computer through an IBM 3270 series communication system. The task and the data collection programs were written in VSAPL.

In the keyboard conditions, subjects communicated with the computer by typing their messages into a 3279 computer terminal. The subjects' keyboard was modified slightly to increase its ease-of-use. First, because the dual function of the ENTER key (i.e. entering information and paging) can be confusing to novice computer users, the paging function was reassigned to the PF1 key. The PF1 key-cap was relabeled to read NEXT PAGE. Subjects were told to press the NEXT PAGE key when their computer screen was full. And second, the ENTER key itself was relabeled to read SEND. The word SEND was used to avoid favoring those subjects who were being shaped to or might model the verb "enter" as opposed to the verb "input".

In the voice conditions, subjects entered their messages by speaking into a miniature headset boom microphone (Shure, model SM10A). The subjects' microphone line was attached to the 3279 terminal via an intermediary box labeled COMPUVOICE III. The subjects were told that COMPUVOICE III translates speech into a form that is understandable to computers. Actually, the COMPUVOICE III box was capable of nothing more than backlighting green and red plexiglass buttons and surrounding a series of lead weights.

Although the subjects in both communication modes believed they were communicating with the computer, they were actually communicating with a human experimenter. The experimenter saw the subjects' inputs on a slaved computer screen in the keyboard mode and heard the subjects' inputs over a set of headphones in the voice mode. She then translated and entered their inputs into the true command-driven applications program on a second terminal. Through a shared variables process, the experimenter's second terminal communicated with the subjects' terminal, and hence the slaved terminal, so all subjects received the program's response to their requests on their own computer screen. The interested reader is referred to Zoltan-Ford (1984) for a more detailed explanation and to Zoltan *et al.* (1982) for a diagram of this paradigm.

Problem solving task

Interactions between the subjects and the computer were confined to an inventory control problem. An inventory task was selected because it is representative of a large class of entry and retrieval programs (e.g. bibliographic searches and management information databases) where the database may not be familiar to its users. The subjects were asked to communicate with the program as if they were employees of an inventory department. Their task was to respond to a series of 30

inter-departmental memos written by other employees in their company. Two of the memos are shown below:

> Billing has just placed an order for 100 packs of mini memo pads. Can we cover that order?
> Reply requested.
>
> Signed Purchasing Dept

> Due to an accident in the warehouse last night (one of our shelves fell down), all but one of our rust ashtrays is now broken. That means that there is *only one rust ashtray* left down here. Sorry for the inconvenience.
> NO REPLY NECESSARY
>
> SIGNED Marty, Warehouse

The memos required the subjects to perform three different types of subtasks: (1) to tell the computer about new inventory; (2) to make changes to existing inventory; and (3) to answer questions about existing inventory. As such, the subjects were creating, manipulating and retrieving inventory files.

To be able to determine if the subjects' language changed as they continued to work with the programs, the 30 sub-tasks were divided into three blocks of 10 subtasks each. To avoid the possibility that one of the three types of actions (creating, manipulating and retrieving files) could become overlearned, each block contained about the same number of task-types. Three of these counterbalanced orders were developed. All three orders were replicated across the 16 treatments. To insure that subjects would adhere to these pre-established orders, the memos were placed in a three-ring binder.

EXPERIMENTAL DESIGN

The experimental design included four completely crossed between-subjects factors, each with two levels. The four independent variables were: (1) the type of communication mode (keyboard or voice); (2) the type of vocabulary used by the computer program (familiar or unfamiliar); (3) the length of the computer program's output phrases (conversational or terse); and (4) the amount of restriction placed on the subjects' language (unrestricted or restricted). Although all four of these variables were chosen to determine the degree to which users' verbal behavior can be altered under various conditions, two of these variables served an even more important purpose. Both the type of vocabulary and the output-length variables defined the language characteristics to which the restricted subjects were shaped.

For the *familiar-vocabulary* subjects, the program displayed words that occur at high frequency in everyday use (e.g. enter, product, change, display). The program displayed words that occur at low frequency for the *unfamiliar-vocabulary* subjects (e.g. input, item, alter, retrieve). The frequency with which these words are used in everyday communications was determined by consulting published word frequency counts (Kucera & Francis, 1967; Carroll, Davies & Richman, 1971).

The programs also differed in the length of the output phrases they used (either conversational or terse). The conversational outputs were complete grammatical sentences whereas the terse outputs included only the verbs and nouns needed to convey the meaning of the message. Wherever possible, the terse outputs did not include pronouns, modal auxiliary verbs, or determiners. In this way the conversa-

tional outputs resembled formal human to human communication and the terse outputs resembled a natural-language version of command languages.

The fourth independent variable was the amount of restriction placed on the subjects' inputs. For the unrestricted subjects the computer performed all the actions they requested regardless of how the inputs were worded. For the restricted subjects the computer only responded to requests that mirrored the computer's language for their vocabularly-output length treatment.

Whenever a restricted subject created an input that did not mimic the program's linguistic characteristics, the computer responded with an error message. Error messages used the vocabulary and phrase-length of the subject's experimental condition. Table 1 shows excerpts from the initial interchanges (where error messages were most likely to occur) of two restricted subjects' interactions with the inventory program.

Messages entered by subjects in the unrestricted condition were analysed to reveal the degree to which people *model* the communication style of a consistently worded natural-language program. Conversely, messages entered by subjects in the restricted condition were analysed to determine the degree to which people's communication style can be *shaped* during interactions with a natural-language program.

DEPENDENT MEASURES

Four verbal dependent variables were used to assess subject performance: messages, mean message length, vocabulary, and output-conforming messages.

Messages: With one exception, messages were the number of unprompted inputs the subject sent to the computer. The one exception involved file manipulation messages. For these subtasks, the program allowed the subjects to use two different strategies. Subjects could convey all necessary information in one message (e.g. "Change unit cost for Pentel pencils to $25·00"). Conversely, subjects could inform the computer that they wanted to manipulate a file (e.g. "Change product") and then could wait for the program to prompt them for the file parameter to be manipulated (e.g. unit cost, item code, etc.). To be able to determine if the two strategies would be used differentially among the 16 conditions, the parameter-prompted messages, when they occurred, were also counted as messages.

Mean message length: The mean message length for each subject was calculated by dividing the number of words used by the number of messages sent to the computer. To derive this measure, the number of words (tokens) used by each subject was first calculated. Words were counted using the guidelines established in earlier telecommunications studies (see, for example, Chapanis, Parrish, Ochsman & Weeks, 1977). The specific guidelines used in the present study are summarized below:

(1) Mispronounced words in the voice mode and misspelled words in the keyboard mode were counted as words.
(2) Abbreviations, such as *invent* for inventory, were counted as words.
(3) Any symbol used in the keyboard mode, such as # for number, were counted as single words.
(4) Words that were inadvertently run together or spliced apart in the keyboard mode were reformed into their appropriate words and counted on the basis of their

TABLE 1

Literal initial interchanges between the inventory program and typical restricted subjects

Conversational-unfamiliar condition

Do you need to input, alter or retrieve an item?

Yes†

I don't understand. Do you need to input, alter or retrieve an item?

I need to know how much we are charging for small waster paper baskets

I don't understand. Do you need to input, alter or retrieve an item?

I need to retrieve an item

What is the item label?

Small waste baskets

Item code	Item label	Unit measure	Unit price	Quantity on-hand
z99309	small waste baskets	each	9·89	300

What do you need to do next?

I wish to input a new item

I don't understand. Do you need to input, alter or retrieve an item?

I need to input an item

What is the item code?

p22898

What is the item label?

Pentel pencils

What is the unit measure of the item?

Box of 12

What is the unit price of the item?

$35·00

What is the quantity on-hand of the item?

60 boxes

The item Pentel pencils has been input.
What do you need to do next?

I need to retrieve an item

What is the item label?

Black ashtrays

Item code	Item label	Unit measure	Unit price	Quantity on-hand
k4372	black ashtrays	each	12·00	500

What do you need to do next?

I need to alter an item

What is the item label?

Black ashtrays

What is it about black ashtrays you need to alter?

TABLE 1 (*continued*)

I need to alter the quantity on-hand to 100

The quantity on-hand for black ashtrays has been altered to 100.
What do you need to do next?

Terse-familiar condition

Enter, change or display product?

Display

Don't understand. Enter, change or display product?

Display product

Product name?

Small waste baskets

Product number	Product name	Unit size	Unit cost	Amount on-hand
z99309	small waste baskets	each	9·89	300

What next?

Enter product

Product number?

c57900

Product name?

Thumb tacks

Unit size?

200 per box

Unit cost?

2·13

Product thumb tacks entered.
What next?

Change product

Product name?

Large waste baskets

Change what about large waste baskets?

Unit cost

Change unit cost for large waste baskets to what?

12·00

Unit cost for large waste baskets changed to 12·00
What next?

Display amount of red ashtrays presently on hand

Don't understand. Enter, change or display product?

Display product

Product name?

Plain white paper

Product number	Product name	Unit size	Unit cost	Amount on-hand
a33358	plain white paper	200 sheets	2·60	155

TABLE 1 (*continued*)

What next?
Change product
Product name?
Plain white paper
Change what about plain white paper?
Product number
Change product number of plain white paper to what?
a33360
Product number of plain white paper changed to a33360.
What next?
Enter product

† The subjects' messages are shown in italics.

revised forms. Thus *displaya* for *display a* and *ware house* for *warehouse* were counted as two words and one word, respectively.
(5) Contractions were split into their constituent words and then counted in their revised forms. *There's,* for example, was split into *there is* and was counted as two words.
(6) False starts and vocal interjections in the voice mode were counted as words. Thus, the utterance *fr from* was counted as two words as was the utterance *uh, yes.*
(7) Because the actual names of the inventory items were provided on the inter-departmental memos, and because they were of no particular interest in the data analysis, all inventory item names used by the subjects were symbolized with a single abbreviation (PN for product name). Thus, whether a subject said *enter imitation leather binders* or *enter paper clips,* both phrases were converted to read *enter PN* and thus were judged to contain only two words.
(8) Any number, whether typed or spoken, was counted as a single word. Thus, both the typed message *change quantity to 21* and the spoken message *change quantity to twenty-one* contain four words. This guideline has not always been adhered to in earlier telecommunications studies. On occasion, for example, the spoken *twenty-one* has been counted as two words while the typed *21* was counted as one word. This method of counting the number of words for numbers was not used, however, because it serves to inflate the number of words used in the voice mode.

Vocabulary: Vocabulary was the total number of different words (types) used by each subject. As with telecommunications studies, all mispronounced, misspelled, and incomplete words were counted as different words from their correct prototypes.

Output-conforming messages: An output-conforming message was a subject-generated message that contained the vocabulary and message length appropriate to the subject's experimental treatment. To permit a comparison among treatments,

output-conforming messages were calculated only for those subject-generated messages used to complete the 30 subtasks (memos). As the reader will see shortly, subjects often generated more messages than the task required.

Additionally, for the restricted subjects, only the first attempted message for each subtask was examined for this dependent variable. For these subjects the computer did not respond to their request until they entered a message that conformed to their program's output. Therefore, if all subject-generated messages were examined for conformity, the number of output-conforming messages for each restricted subject would be the number of subtasks—30. By examining only the first attempted message for each subtask, the number of output-conforming messages in the restricted condition was free to vary.

Attitude measures: Subjects completed both a pre- and a post-experimental attitude questionnaire. The questionnaires consisted of 15 pairs of bipolar adjectives assembled in a semantic differential format (Table 5). The pre- and post-experimental questionnaires differed in two ways. First, the pre-experimental questionnaire instructions asked the subjects to rate computers in general; the post-experimental questionnaire instructions asked the subjects to rate the computer system with which they had just worked. Second, the adjectives were randomly re-ordered both within and among adjective pairs from the pre- to the post-experimental questionnaire.

PROCEDURE

Subjects' sessions were conducted individually. The subjects were told that the purpose of the study was to test some newly developed computer equipment that allows people to communicate in ordinary, everyday English with an inventory control program. Their task as an inventory department employee and the use of the 30 inter-departmental memos was then explained.

Next, subjects completed the pre-experimental attitude questionnaire and then were familiarized with the computer equipment. Keyboard subjects practised typing on the keyboard and voice subjects "trained" the computer to recognize their voice.

After the subject was comfortable with the equipment and procedure, the computer presented a "Letter of Introduction" for the inventory program on the computer screen. Each subject saw the version of the "Letter of Introduction" that corresponded to his or her vocabulary-output length treatment. The letter described the purpose of the natural-language inventory management program. Additionally, it stated that the program occasionally might not comprehend a message because its understanding of natural language was somewhat limited. Under these circumstances, the letter said, the program would display an error message and the subjects then should reword their message and send it to the computer again.

During the initial portion of their interactions with the inventory program, the program guided the subjects. It presented a series of screens that described the three basic types of actions (creating, manipulating and retrieving) that the subject would be able to perform. At this point, the subjects merely read the screens and responded to computer prompts (e.g. Do you want to display the products?). The purpose of this computer-controlled dialogue was two-fold: to gently ease the subjects into the inventory task and to expose them to their program's style of

communication. For the remainder of their sessions, the subjects controlled the interaction by posing the commands and the questions to which the computer responded.

Following their sessions, the subjects completed their post-experimental attitude questionnaires. They then discussed their impressions and choice of input phrases with the experimenter.

Results and discussion

The four verbal measures and the 15 semantic differential attitude scales were each analysed with a four between- and one-within subjects multivariate analysis of variance (MANOVA). For the verbal measures, the within-subjects variable was the three blocks of 10 consecutively ordered sub-tasks. For the attitude scales, the within-subjects variable was the time of the attitude measurement (pre- or post-experimental).

EFFECTS ON VERBAL MEASURES

This MANOVA revealed four significant main effects and five significant interaction effects. None of the significant effects involved the independent variable vocabulary. This finding, or lack of one, shows that for the sets of vocabulary tested, the degree to which people model or can be shaped to a program's communication style is independent of the vocabulary used by the computer.

To determine which verbal measures contributed to the significant multivariate effects, individual four between- and one within-subjects analyses of variance (ANOVAs) were conducted for each measure. Effects that revealed significant differences among groups according to both the MANOVA and ANOVAs are discussed below. Main and interaction effects that are qualified by higher-order interactions are not presented. The presentation of results is organized around the four questions initially raised in the Introduction.

Will people model a program's output?

On average, subjects exposed to conversational outputs generated inputs that contained 60·39% more words per message than did those exposed to terse computer outputs (*Ms* for mean message length = 4·94 and 3·08, respectively), $F(1, 32) = 27{\cdot}05$, $p = <0{\cdot}001$. For comparison, the computer outputs contained, on the average, 5·125 and 2·875 words per message, respectively. Mean message length was also affected by the computer output-length by blocks interaction, $F(2, 64) = 5{\cdot}43$, $p = <0{\cdot}01$. Whereas conversational subjects increased their mean message lengths from the first to the second block of subtasks, $F(2, 46) = 4{\cdot}28$, $p = <0{\cdot}025$, those exposed to terse outputs did not (*Ms* = 4·60 to 5·10 words and 3·16 to 3·06 words, respectively). The lack of a significant interaction between output length and amount of restriction on mean message length shows that people will use the computer's output length as a model for the length of their own inputs.

It is easier for people to model both the length *and* the vocabulary of a terse computer output than of a conversational computer output: the number of output-conforming messages entered by subjects differed as a consequence of the

computer's output length, $F(1, 32) = 6{\cdot}32$, $p = <0{\cdot}025$. Those exposed to terse outputs generated about 40% more output-conforming messages per block of 10 sub-tasks than did those exposed to conversational outputs (*Ms* = 7·15 and 5·11).

Can a program shape people's language?

Averaged across all three sub-task blocks, restricted subjects generated 2·5 times as many output-conforming messages per block of sub-tasks as did unrestricted subjects, $F(1, 32) = 41{\cdot}54$, $p = <0{\cdot}001$. About 90% of restricted users' inputs mirrored the linguistic characteristics used by their programs while only 35% of unrestricted users' inputs did (*Ms* = 8·75 and 3·51 per block).

As shown in Table 2, the number of output-conforming messages was also affected by the restriction × output length × blocks interaction, $F(2, 64) = 10{\cdot}95$, $p = <0{\cdot}001$. An examination of this table reveals two noteworthy findings. First, the only length-restriction treatment combination that did not lead subjects to generate more output-conforming messages as they progressed from the first to the second block of sub-tasks was the conversational output with unrestricted inputs, $F(2, 22) = 0{\cdot}66$, $p = >0{\cdot}05$. Second, subjects exposed to the terse outputs were easier to shape than were those exposed to the conversational outputs: although the restricted subjects exposed to terse and to conversational outputs did not differ at blocks II and III, those exposed to terse outputs entered more output-conforming messages during block I than did those exposed to conversational outputs, $F(1, 22) = 8{\cdot}21$, $p = <0{\cdot}01$.

The three-way interaction among restriction, output-length, and blocks also affected the number of vocabulary (unique words) the subjects used, $F(2, 64) = 3{\cdot}66$, $p = <0{\cdot}05$. As shown in Table 3, all four groups decreased the number of new vocabulary they used after the first block. During the first block, unrestricted subjects used about 21 unique words, regardless of the length of the computer's output. The number of unique words used by restricted users, however, differed during the first block, $F(1, 22) = 9{\cdot}69$, $p = <0{\cdot}01$. At the beginning of their task, restricted subjects exposed to conversational outputs generated about 1·5 times more unique words than those restricted to terse outputs.

TABLE 2

Mean number of output-conforming messages† entered by restricted and unrestricted subjects as a function of computer output length and sub-task block

	Sub-task block		
Condition	1	2	3
Restricted			
Terse	8·17	9·67	9·67
Conversational	6·50	9·00	9·50
Unrestricted			
Terse	4·42	5·25	5·75
Conversational	2·17	1·75	1·75

† Maximum possible number of output-conforming messages per block is 10.

TABLE 3
Mean new vocabulary introduced by restricted and unrestricted subjects over the three sub-task blocks as a function of computer output length

	Sub-task block		
Condition	1	2	3
Restricted			
Terse	16·58	1·08	2·08
Conversational	24·92	1·32	2·92
Unrestricted			
Terse	21·42	2·67	4·08
Conversational	20·50	5·25	6·08

The process of shaping the restricted users came at a cost—restricted subjects sent more messages to the computer during the first 10 subtasks than during later blocks, $F(2, 46) = 38{\cdot}25$, $p = <0{\cdot}001$, and more than unrestricted subjects did during their first block, $F(1, 46) = 32{\cdot}19$, $p = <0{\cdot}001$. Inspection of the subjects' communication protocols reveals that about 18% of the restricted subjects' messages can be attributed to a learning process. In other words, collapsing over the three subtask blocks, restricted subjects sent an average of 10·54 improperly formatted messages to the computer. This learning process accounts for 80% of the disparity in messages between the two restriction conditions. The remainder of the restricted subjects' additional messages are due to differences in the approaches subjects took to manipulate files, χ^2 $(2, N = 48) = 6{\cdot}95$, $p = <0{\cdot}05$. More often than not, restricted subjects told the computer they wanted to manipulate a file. Then they waited for the program to prompt them for the file parameter to be manipulated (quantity on-hand, unit price, etc.). Thus many restricted subjects used two messages to complete each manipulation sub-task. Unrestricted subjects, on the other hand, were more inclined to send the computer both pieces of information in one message; the need to manipulate and the necessary information. To illustrate this difference, portions of typical restricted and unrestricted sessions are shown in Table 4.

LINGUISTIC DIFFERENCES WITHIN RESTRICTION CONDITIONS

A three between- and one-within subjects ANOVA was performed on the unrestricted subjects' messages using a relaxed output-conforming message criterion. To be counted as an output-conforming message under the relaxed criterion, the message had only to include the correct verb (e.g. enter, input, alter, retrieve). No significant effects, main or interaction, were found.

About 78% of the unrestricted users' inputs contained the verb found in the output-conforming messages. For the phrase "*enter* product", subjects frequently gave "please enter product", "enter new product into inventory", and "enter". Therefore, what prevented 27% of the unrestricted subjects inputs in the terse condition and 59% in the conversational condition from being counted as output-conforming messages, was the inclusion of phatic expressions (e.g. "please", "into inventory", and "in warehouse"); the exclusion of portions of the intended phrase

TABLE 4

Literal interchanges between the inventory program and typical restricted and unrestricted subjects: file manipulation strategies

Restricted condition
What next?†
Change product
Product name?
Gregg steno books
Change what about Gregg steno books?
Amount on hand
Change amount on-hand for Gregg steno books to what?
Zero
Amount on-hand for Gregg steno books changed to 0·0 What next?
Unrestricted condition
What do you want to do next?
I want to change the amount on-hand for Gregg steno books to 0·0
The amount on-hand for Gregg steno books has been changed to 0·0 What do you want to do next?

† The subjects' messages are shown in italics. Both excerpts use familiar vocabulary. The restricted interchanges are from the terse condition, the unrestricted from the conversational.

(e.g. "enter" instead of "enter product"); or the substitution of semantic correlates of other portions of the intended phrase ("I *would like* to enter a product" or "I *wish* to enter a product" instead of "I *want* to enter a product").

The remaining 22% of the unrestricted subjects' messages contained synonyms of the intended verbs. Most common among these was the verb "to add", which was used in place of the file creation verbs (i.e. enter and input). Other substitutions involved "print" and "see" for "retrieve", and "revise" and "adjust" for "alter".

The types of messages originally generated by the restricted subjects were like those of the unrestricted subjects. Many messages included phatic expressions and semantic correlates or lacked required vocabulary. As opposed to the unrestricted subjects, however, the restricted subjects quickly changed their inputs to be like the program's output. They did so to be able to continue with their task. Nonetheless, shaping did not occur at the same rate across the restricted conditions, as indicated by the results of an ANOVA performed on the number of *non*-output-conforming messages entered by the restricted subjects. The number of non-output-conforming messages that restricted subjects entered is the number of false starts they experienced before they finally created an output-conforming message. Only one effect was significant—computer output-length, $F(1, 16) = 10{\cdot}40$, $p = <0{\cdot}01$. In general, subjects restricted to conversational inputs entered 2·78 times more non-output-conforming messages (false starts) than did subjects restricted to terse inputs.

The difference in difficulty between shaping conversational and terse subjects, as shown by the non output-conforming messages and message measures, and the

difference in degree of modeling between conversational and terse subjects, as shown by the output-conforming messages, is due to the increased likelihood of errors in inputs containing many, as opposed to only a few, words. Conversational subjects, unlike terse subjects, generated "lead in" phrases (e.g. "I want to . . ."). As many of the conversational subjects later commented, they more frequently use the verb phrase "would like" rather than the verbs "want" (familiar) or "need" (unfamiliar) in their interpersonal communications. They then transferred this preference to their interactions with the computer. As a consequence, conversational subjects were less likely to generate output-conforming messages initially in the restricted conditions and throughout their sessions in the unrestricted conditions.

Are there mode differences?

Differences between voice and keyboard subjects were seen on two dependent variables (messages and mean message length) through the mode of communication × blocks of sub-tasks interaction. The mode × blocks interaction effect on messages, $F(2, 64) = 8{\cdot}10$, $p = <0{\cdot}001$, is caused by the different number of messages sent by voice *vs* keyboard subjects during the second and third blocks of subtasks. During these later blocks, voice subjects sent about 38% more messages (*Ms* = 16·79 and 18·50) to the computer than did keyboard subjects (*Ms* = 12·88 and 12·66), $F(1, 46) = 11{\cdot}88$ and 12·44, $p = <0{\cdot}025$ for blocks 2 and 3. The additional voice messages are due to two primary sources: retrieving and manipulating inventory files. During the second and third blocks of subtasks, voice subjects retrieved nearly two times as many inventory files as did keyboard subjects, even though the sub-tasks did not require them to do so, $F(1, 46) = 6{\cdot}99$ and 5·56, $p = <0{\cdot}025$.

Two explanations for this difference seem apparent: (1) as compared with keyboard subjects, voice subjects felt less confident that their messages were understood correctly by the computer and thus retrieved the information they had previously entered to check its accuracy; or (2) because speaking requires less effort than does typing, voice subjects checked the computer's accuracy more than keyboard subjects, because it was easier to do so. Data from the post-experimental interviews supports the latter hypothesis. Both groups of subjects said they initially were uncertain that the computer understood them. Keyboard subjects, however, were satisfied that the computer did understand after they retrieved a few of the files they just had manipulated or created. Voice subjects, although similarly convinced, did not find this strategy required additional effort and so continued to retrieve files.

Beginning with the second block, voice subjects also sent more manipulation messages to the computer than did keyboard subjects. Voice subjects most often used the two-message approach; keyboard subjects most often used the one-message approach, χ^2 (2, $N = 48$) = 10·05, $p = <0{\cdot}01$.

The mode by block interaction effect on mean message length follows from the effect on messages noted above. While those who communicated by voice did not change, those who communicated by keyboard increased their mean message lengths from the first to the second block of sub-tasks, $F(2, 46) = 6{\cdot}73$, $p = <0{\cdot}005$. The increased message lengths for the keyboard subjects is explained by their manipulation strategy described above. During the first block, keyboard subjects used the same two-message approach as did voice subjects. By the second block, they had learned to enter a manipulation request in one somewhat longer message.

How are users' attitudes affected?

The attitude-measures MANOVA revealed one significant main, and five significant interaction effects. All six significant effects involved within-subject differences. Four between- and one within-subjects ANOVAs were performed on all 15 adjective scales to determine which scales contributed to the MANOVA effects.

Only two of the effects are discussed below. The four remaining interaction effects occurred on attitude scales where the subjects differed on their pre-experimental responses but not on their post-experimental responses.

Attitude measurements. The subjects' attitudes toward computers changed on 12 of the 15 scales following their interactions with the program ($p = <0{\cdot}001$ for each scale). Mean pre- and post-experimental responses to the 15 scales are shown in Table 5.

The subjects' pre- and post-experimental attitudes can be summarized in two statements. First, if their sessions with the computer had an effect on their responses to the attitude scales, it caused their attitudes to become more positive. Second, if their sessions did not cause them to change their responses, their attitudes remained at the same positive level as they had before their sessions.

Amount of restriction—attitude measurement interaction: Subjects' responses to two of the attitude scales changed as a function of the amount of restriction placed

TABLE 5

Mean responses to pre- and post-experimental attitude questionnaire†

	Time of measurement	
Adjective pairs	Pre-	Post-
Personal–impersonal	4·52	3·35‡
Simple–complicated	4·83	2·29‡
Helpful–hindering	1·73	1·69
Systematic–random	1·82	1·76
Easy–difficult	4·27	1·96‡
Forgiving–unforgiving	4·77	2·81‡
Obedient–bossy	3·29	1·96‡
Cooperative–obstinate	2·96	1·75‡
Unthreatening–threatening	3·29	1·92‡
Intelligent–simple-minded	2·81	2·79
Pleasing–disgusting	2·66	2·00‡
Flexible–inflexible	3·42	2·56‡
Satisfying–frustrating	3·56	2·06‡
Calming–anxiety–provoking	4·71	3·08‡
Obliging–demanding	3·79	2·48‡

† Attitude scales are in the order used for pre-experimental measurement. The order of adjectives within bipolar pairs has been rearranged: the more positive adjective always appears on the left, the more negative adjective appears on the right. Mean values could range from 1 (extreme agreement with the first, or more positive, adjective) to 7 (extreme agreement with the second, or more negative, adjective). A mean of 4 indicates a neutral response.

‡$p = <0{\cdot}001$.

on their communications. First, the unrestricted subjects felt computers were more *flexible* after their interactive sessions, $F(1, 23) = 16{\cdot}79$, $p < 0{\cdot}001$, but the restricted subjects did not. Such a finding confirms that the restricted subjects were aware that the program's understanding was limited and that the unrestricted subjects discovered that the program could interpret messages with simple spelling errors and variations in input language.

Second, restricted subjects expressed the same *satisfaction* with computers after as they had before their sessions. Unrestricted subjects, in contrast, felt more satisfied with computers after their sessions, $F(1, 23) = 33{\cdot}61$, $p = <0{\cdot}001$. Although the lack of flexibility does not result in reduced satisfaction, it is a necessary component to occasional users' increased satisfaction with computers.

POST-SESSION INTERVIEWS

Regardless of condition, subjects' reactions following their sessions were overwhelmingly positive. Subjects freely commented on the ease and enjoyment of using the system.

When asked what types of words they used during their interactions, most subjects said they used the same words as the program did. They explained that it was easy to do and they were fairly sure the program would understand words that it used. A few subjects, whether restricted or not, stated this explanation in a slightly different way: They used the program's words because they were "afraid that it wouldn't work otherwise". Unrestricted subjects commented, however, that they were pleasantly surprised when the program understood words and phrases that it had not used.

Restricted subjects did not feel hampered by the experimental conditions. They offered two related explanations for this reaction: (1) because computers do not have the vocabulary that people do, programs must require at least some consistency; and (2) as in person to person interactions, computer users must learn how to communicate with their interactive partner (the computer).

Nearly all subjects were surprised to find the bipolar adjective pair "personal-impersonal" on the pre-experimental questionnaire. Following their sessions, however, the subjects commented about three aspects of the program's personal nature. First, restricted subjects said the program's error messages did not insult or threaten them, but rather politely told them it did not understand and asked them to rephrase their message. Second, regardless of condition, subjects liked the program's use of *please* and *thank you* in messages to remind them to turn the page and in the message at the end of their sessions, respectively. Third, the subjects remarked that the wording of the program's responses created a feeling of personal interchange, like that of a normal conversation between people.

Conclusions

The purpose of this study was to examine the possibility of modeling and shaping effects in natural-language interactions with computers. The specific questions of interest were listed in the Statement of Purpose section. They are answered in turn below.

First, will people model the linguistic characteristics of a program's output? Users of natural-language programs will model the length of the program's output regardless of the length it uses (conversational or terse). They will do so, in addition, regardless of the vocabulary used by the program (familiar or unfamiliar) or of the mode of input provided to them (voice or keyboard). Furthermore, it is easier for people to model both the length and the vocabulary of a terse computer output than of a conversational one. In neither case, however, is the modeling perfect. Even in the terse condition, only an average of 51·4% of the subjects' inputs mirrored the program's outputs perfectly.

Nonetheless, the messages sent by subjects in the terse and conversational unrestricted groups were often not that different from the output-conforming messages required of the restricted subjects. About 78% of their inputs contained the verb found in the output-conforming messages.

Second, shaping is more effective in reducing the variability in people's language than is relying on them to model. As with modeling, it is easier to shape people to terse than to conversational outputs. This difference occurs because terse outputs contain less vocabulary than do conversational outputs. However, the increased difficulty of shaping users to conversational inputs is shortlived. If restricted users create unacceptable messages, they do so almost entirely in the early, or learning stages of their interactions.

Third, there is no difference in modeling or shaping effects between spoken and typed inputs. However, differences other than those directly related to modeling and shaping did occur between the two modes of communication. When people speak to a computer they are more likely to take a conservative approach to file manipulation than when they type their inputs. In so doing, voice users allow the computer to prompt them for details of their manipulation requests. Keyboard users, conversely, are more likely to input all necessary manipulation information in one message. This same manipulation strategy difference is seen occasionally between restricted and unrestricted users. People whose language is restricted tend to let the computer prompt them for information whereas people whose language is not restricted do not.

Those who use voice inputs also differ from those who use keyboard inputs in the number of retrieval requests they make. People who use voice inputs often retrieve files before and after they manipulate or create these files, regardless of their certainty that the computer has understood their requests. People who use keyboard inputs do this only during their initial interactions with the program. After they have assured themselves that the program understands their requests, those who use keyboard inputs no longer retrieve such files.

Lastly, occasional users of computer systems respond in an extremely positive manner to natural-language input systems like the one tested here. In not one instance did subjects' attitudes toward computers become more negative following their interactions with the system. Rather, subjects' reactions became almost uniformly more positive as shown by significant increases in reactions to such descriptors as personal, simple, pleasing, calming, easy and obliging. The primary exception to this finding is that unrestricted subjects felt more satisfied with computers following their interactions with the natural-language program, but restricted subjects did not. Restricted subjects remained at the same level of

satisfaction after as they had before their sessions with the program (slightly satisfied).

The conclusion to be drawn from this research is that variability of expression in natural-language interactions with computers need no longer present a problem to designers of such systems. Recognition rates of natural-language processors will increase if designers implement the following three criteria:

(1) Provide a consistently worded program output: users will model it;
(2) Design the program to communicate with tersely phrased outputs of the form *verb–noun*: users will model this format more than they will conversational outputs of the form *pronoun-modal auxiliary-verb-determiner-noun*;
(3) Include non-threatening error messages that reiterate the vocabulary and phrases that the processor can understand: if the program cannot understand their initial requests, users will alter their vocabulary and phrase structures to be like those provided in the error messages.

Users of natural-language systems designed according to these guidelines will feel less anxious about computers and will consider computers easy to use and personal. The key to increased satisfaction with computers, however, lies in the processors' ability to handle variability: the greater the natural-language processor's ability to accept misspellings and personal preferences in word choice and phrase structure, the more satisfied its users will be with computers.

The author acknowledges and thanks GTE Laboratories, Inc. of Waltham, Massachusetts for providing the funding and equipment for this research.

References

BALL, R. S. (1952). Reinforcement conditioning of verbal behavior by verbal and non-verbal stimuli in a situation resembling a clinical interview. (PhD dissertation, Indiana University, 1952). *Dissertation Abstracts,* **12,** 131.

BECKER, J. (1975). The phrasal lexicon. In R. SCHANK & B. NASH-WEBBER, Eds. *Theoretical Issues in Natural Language Processing,* pp. 38–41. Cambridge, MA: Association for Computational Linguistics Workshop, (ACL).

BLACK, J. B. & MORAN, T. P. (1982). Learning and remembering command names. In *Human Factors in Computer Systems Proceedings,* pp. 8–11. Gaithersburg, MD: US Department of Commerce.

CARROLL, J. B., DAVIES, P. & RICHMAN, B. (1971). *Word Frequency Book.* New York: American Heritage Publication.

CASSOTTA, L., FELDSTEIN, S. & JAFFE, J. (1968). *The Stability and Modifiability of Individual Vocal Characteristics in Stress and Non-stress Interviews,* Research Bulletin No. 2. New York: The William Alanson White Institute.

CHAPANIS, A., PARRISH, R. N., OCHSMAN, R. B. & WEEKS, G. D. (1977). Studies in interactive communication: II. The effects of four communication modes on the linguistic performance of teams during cooperative problem solving. *Human Factors,* **19,** 101–126.

DANZINGER, K. (1976). *Interpersonal Communication.* New York: Pergamon Press.

FORD, W. R. (1981). Natural-language processing by computer—a new approach (PhD Dissertation, The Johns Hopkins University, 1981). *Dissertations Abstracts International,* **42,** 808B.

GREENSPOON, J. (1955). The reinforcing effect of two spoken sounds on the frequency of two responses. *American Journal of Psychology,* **68,** 409–416.

HARRIS, L. R. (1977). User oriented data base query with the ROBOT natural language query system. *International Journal of Man–Machine Studies,* **9,** 697–713.

HILDUM, D. C. & BROWN, R. W. (1956). Verbal reinforcement and interviewer bias. *Journal of Abnormal and Social Psychology,* **53,** 108–111.

JAFFE, J. (1964). Verbal behavior analysis in psychiatric interviews with the aid of digital computers. In D. McK. RIOCH & E. A. WEINSTEIN, Eds. *Disorders of Communication,* pp. 389–399. Baltimore, MD: Williams and Wilkins.

KELLY, M. J. & CHAPANIS, A. (1977). Limited vocabulary natural language dialogue. *International Journal of Man–Machine Studies,* **9,** 479–501.

KUCERA, H. & FRANCIS, W. N. (1967). *Computational Analysis of Present-day American English.* Providence, RI: Brown University Press.

MALHOTRA, A. (1975). *Design Criteria for a Knowledge-based English Language System for Management: An Experimental Analysis,* Project MAC TR–146. Cambridge, MA: Massachusetts Institute of Technology.

MANDLER, G. & KAPLAN, W. K. (1956). Subjective evaluation and reinforcing effect of a verbal stimulus. *Science,* **124,** September, 582–583.

MATARAZZO, J. D. (1962). Prescribed behavior therapy: Suggestions from non-content interview research. In A. J. BACHRACH, Ed. *Experimental Foundations of Clinical Psychology,* pp. 471–509. New York: Basic Books.

MATARAZZO, J. D., HESS, H. F. & SASLOW, G. (1962). Frequency and duration characteristics of speech and silence behavior during interviews. *Journal of Clinical Psychology,* **18,** 416–426.

MATARAZZO, J. D., SASLOW, G., MATARAZZO, R. G. & PHILLIPS, J. S. (1958). Stability and modifiability of personality patterns during a standardized interview. In P. A. Hoch & J. Zubin, Eds. *Psychopathology of Communication,* pp. 98–125. New York: Grune & Stratton.

MATARAZZO, J. D., WEITMAN, M., SASLOW, G. & WIENS, A. N. (1963). Interviewer influence on duration of interviewee speech. *Journal of Verbal Learning and Verbal Behavior,* **1,** 451–458.

MCNAIR, D. M. (1957). Reinforcement of verbal behavior. *Journal of Experimental Psychology,* **53,** 40–46.

MOCK, J. F. (1962). The influence of verbal and behavioral cues of a listener on the verbal productions of the speaker. (PhD dissertation, University of Kentucky, 1957). *Dissertation Abstracts,* **23,** 312.

NUTHMAN, A. M. (1957). Conditioning of a response class on a personality test. *Journal of Abnormal and Social Psychology,* **54,** 19–23.

PETRICK, S. R. (1976). On natural-language based computer systems. *IBM Journal of Research and Development,* **20,** 314–325.

RAY, M. L. & WEBB, E. J. (1966). Speech duration effects in the Kennedy news conferences. *Science,* **153,** 899–901.

SARASON, B. R. (1957). The effects of verbally conditioned response classes on post-conditioning tasks (PhD dissertation, Indiana University, 1956). *Dissertation Abstracts,* **17,** 679–680.

TAFFEL, C. (1955). Anxiety and the conditioning of verbal behavior. *Journal of Abnormal and Social Psychology,* **51,** 496–501.

THOMPSON, B. H. (1980). Linguistic analysis of natural language communication with computers. In *Proceedings of the Eighth International Conference on Computational Linguistics (COLING '80),* Tokyo.

ZOLTAN, E., WEEKS, G. D. & FORD, W. R. (1982). Natural-language communication with computers: A comparison of voice and keyboard inputs. In G. JOHANNSEN & J. E. RIJNSDORP, Eds. *Analysis, Design and Evaluation of Man–Machine Systems,* pp. 287–292. Dusseldorf, Germany: Verein Deutscher Ingenieure.

ZOLTAN-FORD, E. (1984). Language shaping and modeling in natural-language interactions with computers. (PhD dissertation, The Johns Hopkins University, 1983). *Dissertation Abstracts International,* **44,** 3563B.