LINGUISTIC ANALYSIS OF NATURAL LANGUAGE COMMUNICATION WITH COMPUTERS

Dr. Bozena Henisz Thompson
California Institute of Technology
Pasadena, California, USA

Summary

Interaction with computers in natural language requires a language that is flexible and suited to the task. This study of natural dialogue was undertaken to reveal those characteristics which can make computer English more natural. Experiments were made in three modes of communication: face-to-face, terminal-to-terminal and human-to-computer, involving over 80 subjects, over 80,000 words and over 50 hours. They showed some striking similarities, especially in sentence length and proportion of words in sentences. The three modes also share the use of fragments, typical of dialogue. Detailed statistical analysis and comparisons are given. The nature and relative frequency of fragments, which have been classified into twelve categories, is shown in all modes. Special characteristics of the face-to-face mode are due largely to these fragments (which include phatics employed to keep the channel of communication open). Special characteristics of the computational mode include other fragments, namely definitions, which are absent from other modes. Inclusion of fragments in computational grammar is considered a major factor in improving computer naturalness.

The majority of experiments involved a real life task of loading Navy cargo ships. The peculiarities of face-to-face mode were similar in this task to results of earlier experiments involving another task. It was found that in task oriented situations the syntax of interactions is influenced in all modes by this context in the direction of simplification, resulting in short sentences (about 7 words long). Users seek to maximize efficiency in solving the problem. When given a chance, in the computational mode, to utilize special devices facilitating the solution of the problem, they all resort to them.

Analyses of the special characteristics of the computational mode, including the analysis of the subjects' errors, provide guidance for the improvement of the habitability of such systems. The availability of the REL System, a high performance natural language system, made the experiments possible and meaningful. The indicated improvements in habitability are now being embodied in the POL (Problem Oriented Language) System, a successor to REL.

I. Introduction

The research reported on is part of a larger project aimed at improving the interaction of humans with computers in a language that is natural for the user. In real life applications of computers the language is natural in a very specific sense, since it is constrained by the linguistic and situational context and subject to the inevitable restrictions of the computational grammar and the general requirements of this mode of interaction. However if computational interaction is to be natural, forms of language which are natural in normal dialogue as well as those particularly suited to the application should be available to the user. A very important requirement is that the means of communication be flexible, that the user should be able to modify the language so as to best serve the solution of the problem. Another issue is to what extent the computer should act as a natural party to the interaction. Naturalness of human-computer interaction is often refered to as a system's habitability.

This research was undertaken upon the belief that investigation of human dialogue (both spoken and written) and analysis of human-computer interaction is essential to determine how good habitability can be achieved. Initial research was done on human-to-human dialogue, both in the face-to-face mode in totally free (but voice only) interaction and in the written mode where the dialogue was via computer terminals linked to a computer system, but where the interaction was in unrestricted English. Initial research involved the solution of a relatively simple though quite realistic problem. It confirmed some expected differences between the two modes of communication, but also revealed some surprising similarities. An extremely important result and one that proved particularly challenging to obtain was the identification and definition of structures other than sentences used in natural unrestricted communication. These were finally reduced to about a dozen categories. The next stage of research involved a real-life task and the data was the same for three modes: face-to-face, terminal-to-terminal, and human-to-computer. Results for the first two modes were closely comparable to the previous results and were also compared with results from the computational mode. Again, what is more striking and worthy of interest are the similarities rather than the differences. Some of the major similarities are in sentence length, percentage of words in sentences (as against fragments), number of sentences (for terminal-to-terminal and human-computer mode), very high number of sentences containing be-verbs, and low number of sentences containing relative pronouns.

In this paper, the focus is on (1) the comparison of statistics obtained for the three

modes; (2) the nature and relative frequency of fragments and their implications for computational habitability; and (3) detailed discussion of the characteristics of the computational interactions.

The research involved over 100 subjects in years 1975, 1977 and 1979/80. The subjects were predominately undergraduate and graduate students at Caltech. This work resulted in an enormous amount of data, requiring a great deal of time for analysis. Since each protocol was scored by at least two people (and usually more), averaging out the scores was also time consuming. Total time spent by subjects in experiments was over 50 hours, which yielded for final comparisons 20 face-to-face protocols, 11 terminal-to-terminal, and 21 human-to-computer, containing over 80,000 words. Protocols of over 20 subjects in face-to-face and terminal-to-terminal mode were analyzed for categories and partial statistics, and thus not included in the final results.

The main thesis of this paper is that in problem solving situations ordinary conversation and human-computer conversation in a system that allows relative natural language, share several important features, and that we can improve computer habitability by learning about the nature of ordinary conversation, which exhibits rather well defined and identifiable structural patterns.

II. Early Experiments

In an interesting paper on natural human dialogue in a problem solving situation[1], it was noted on the basis of extensive experiments that "people do not naturally speak in sentences" and that in general great unruliness characterizes interactive communication, whether spoken or written. At first sight of the protocols, one tends to confirm the impression. But a closer look both at the same protocols and at the results of analysis cited, as well as some informal observation of other conversations, and the reflection that communication would hardly be achievable in such an absence of rules, led me to a hypothesis that there is considerable order in natural conversation. I designed experiments in the summar of 1975 and they were conducted in the fall with the assistance of students in a course in Sociolinguistics at Caltech. Additional experiments were conducted in 1977. These experiments are discussed in some detail since they provided guidelines for future research. They differ from the later experiments in the fact that subjects used as much time as was needed for the solution of the experiment, while in the later ones an arbitrary cut-off was imposed. The problem was that of locating the nearest doctor to a patient's address, given a map of Pasadena and a selected list of doctors. Each experiment involved two subjects, one being given the map of Pasadena with the patient's address marked on it (3 different locations were used, but only one in a given experiment) and the other the list of doctors. In the face-to-face mode, the conversations were tape-recorded and transcribed. Subjects were free to communicate by voice but were not allowed to look at each others' materials. Typically, they were seated at the ends of a fairly large table with the tape-recorder between them, and the experimentor in the room. The experimentor provided the materials and instructions, and answered some initial questions only. In the terminal- to-terminal mode, the subjects were in separate rooms and the protocols were recorded and merged with a computer program. The subjects were free to communicate in ordinary English but had to observe some minimal typographical conventions (such as sending the message in by using two keys simultaneously). The role of the experimentor was the same, but divided between the subjects and occasionally offering assistance with computational requirements. The subjects were fully aware that they were conversing with a human counterpart (in this, these experiments differed from Bill Martin's.[2] For the purposes of this paper, the results of 12 face-to-face experiments and 7 terminal-to-terminal are used.

The problems of analysis were severe, but the results gratifying. As noted by other investigators, conversational English has a great deal of characteristics which call for an approach very different from the analysis of well formed single sentences. What is obvious is that there are many strings which are not sentences but "incomplete" or "unfinished" or ungrammatical in a variety of ways. But the difficulty of even deciding what a sentence is has also been noted (3): "...the sentence is not, strictly speaking, a unit in oral discourse. One can see texts in which long sequences of clauses linked by 'and then...' occur. Are these separate sentences or one sentence?" After considerable reflection and search for guidance from the literature, a rather conventional notion of sentence was used, with the requirement that it contain a NP and a VP and that it be within the confines of a single message (a message being the utterance(s) of one speaker). Semantic considerations (admittedly often inevitably intuitive) were used to determine single or multiple sentencehood. Coordinating conjunctions and sequences such as "and then", pauses or phatic units (these being defined as any strings keeping the channels of communication open) often signalled separate sentences. Words such as "because" or "if" tied strings into a single sentence. An additional semantic requirement (again admittedly vague) was that a sentence could stand alone as a unit and make sense. These criteria worked quite well as evidenced by the counts made by different scorers of the same protocols.

An interesting category of sentences that emerged were the so-called transposed sentences, e.g., "The two small streets there might be

doctors on.", "Conwire, I also have.", "4, is it?", "That's the tallest thing we got, special weapon?", "Length by width, does it matter?". Although they are infrequent, such sentences contribute considerably to the distinctive impression made by ordinary conversation. Due to their low frequency, only a partial analysis of these was made. (In four protocols, they amounted to about 2.5% of the total of sentences.) The first three examples above show only word rearrangement, but the others contain pronouns substituting for the transposed NP, in one case preceding the NP, in the other following it.

Some problems were encountered in the consideration of what was a word in numbers, abbreviations, alphanumeric strings. In general numbers and abbreviations were considered one word; in alphanumeric strings, a number was one word, and a character string another. Phatics such as "uh", "um", "uhuh", as well as "okay", even when abbreviated to "O.K.", were considered one word, and as multiple words when obviously so, as in "you know", "I see". Differences in exact word counts were not large enough to be significant, and often coincided surprisingly well.

The most severe problem was, naturally, the definition of fragments. Even though a fairly clear classification was formulated at the end of the analysis of the 1977 experiments, fragments and phatics are discussed in conjunction with the 1979/80 results. Their definition was refined and some categories were reformulated, and comparisons are made with computational protocols. The role and desirability of these fragments in natural conversation is also discussed then.

The most significant results of the early experiments are summed up in Table 1.

TABLE 1
Striking Characteristics
of Face-to-Face Conversations

Short sentences (average length: 7 words, 90% under 10 words long)
~70% of words are in sentences
~20% of words are in fragments
~10% of words are phatics
Fragments help convey more information
Phatics keep attention and kill silence
Minimal insertions complete fragments

Table 2 illustrates the main differences between the face-to-face (F-F) and terminal-to-terminal (T-T) modes.

TABLE 2

	F-F	T-T
Time	X	over 2X
Number of words	X	~3X
Words/minute	~65	~10
Phatics	~10%	~5%
Percent of messages containing fragments	~70%	~50%
Words/sentence	~7	~7
Words/message	~10	~10
Total words in sentences	close to 70%	slightly over 70%

III. Three Modes of Communication: A Comparison

1. The Experimental Setting

The setting in the 1979/80 experiments in the F-F and T-T modes was similar to that described in Section II, but with major differences in the overall design of the experiments. First, three modes of communication were used, the third being human-to-computer. The task was a real life task of loading cargo onto a ship, the data being from the real environment of loading U.S. Navy ships by a group located in San Diego, California. In the first two modes, one subject was provided with a list of cargo items to be loaded (along with their quantities) and a list of decks, their sizes, and their primary uses. The other subject was given a list of the sizes of the cargo items. The subjects were instructed to obey space and other limitations (e.g., hatch size) and restrictions as to what cargo could be stowed on what decks. There was a time limit of one hour in both modes. The task of transcribing F-F recordings was very laborious due primarily to the specific jargon and numerous abbreviations in the data. In the T-T mode, the protocols were obtained automatically.

For the human-computer mode, the REL System was used.[4-6] This system, developed by our project, provides the means for communicating with a large data base in a limited but useful style of natural English, described in detail in.[6] The response times to user queries are quite reasonably short so that natural interaction is possible. Requests which are not understood are diagnosed extremely quickly, thus encouraging the user to try alternate ways of phrasing. This technique was indeed employed frequently, as discussed in Section IV.

It has always been the REL System's philosophy that naturalnesss of a language is obtained in two primary ways: task-specificity and flexibility for modifications. Task-specifity can be achieved only by actual study of the users' needs (and, obviously, by incorporating their data in the system). The capabilities of REL English have already been extended to make the language more natural for this specific task, notably by developing a

prompting "load sequence" (and "offload sequence") in which the computer elicits the information from the user, and offers clarification if the prompt is not clear. This device was used extensively by the subjects, but its description is left out due to space limitations.

The other major ingredient of naturalness is enabling the user to suit the language to the task by incorporating his specific knowledge and jargon. To do this, the user must be able to extend the language through definitions and make other modifications. This, also, was done by the subjects and is discussed in Section IV.

The experimental setting was obviously very different. One subject at a time was assigned the task. No precise time limit was set, but most subjects were given two hour time slots, some of which was spent in initializing the computational session. The subject's session on the average lasted one and a half hours. The subjects were given a list of the cargo items to be loaded and the number of each, as well as the primary uses of the decks. They were instructed that they should attend to fitting the cargo through hatch sizes and to keep track of space loaded. All the pertinent data about the cargo and ships was in the computer. The subjects were also given a short manual on the loading of ships, with examples of how to use the system and English, including arithmetic, definitions and load sequences. The experimentor helped the subject get started and assisted in case of computational problems in about half of the cases, others working alone. Although the subjects were instructed to read the manual before commencing experiments, the analysis of protocols showed that few had actually familiarized themselves with the system.

2. The Structure of Face-to-Face Dialogues

Some working definitions need to be stated here. Messages and sentences were discussed in Section II. Fragments are all of the dialogue material that is not in sentences, and Phatics, which constitute a big subgroup of fragments, are all strings which serve a variety of functions which may all be characterized as keeping the channel of communication open (including expressions of emotions to the other subject and the computer).

A page from a dialogue in Figure 1 illustrates some of the problems in analysis, and gives an idea of some of the categories of fragments, since it contains a rather large number of them. The categories are defined after the discussion of the page. Abbreviations are:

S = sentence	ADD = added information
P = phatic	SELF = talking to oneself
C = connector	TQ = terse question
TR = terse reply	TI = terse information
FS = false start	INT = interrupted
E = echo	TRUN = truncated
TRANS = transposed sentence (discussed in Section II)	

FIGURE 1

1	A	There are like five categories right there. Ammunition, pyrotechnics, special weapons, vehicles and things on pallets. [ADD]
2	B	Yes, let's run through this list here. [P]
3	A	Okay. [P]
4	B	CHG demo. [TQ/TI]
5	A	CHG. Okay, I have one CHG demo here on page 2. [E,P]
6	B	And that page consists of? [C,TQ]
7	A	It's a 32... [INT/TRUN]
8	B	I, I just want to, want to know what type. [FS]
9	A	Oh, that's an ammunition. [P]
10	B	It's an ammunition. [E]
11	A	Yeah. [P/TR]
12	B	Uh hum, and some conwire? [P, C, TQ]
13	A	Conwire, I also have. That's... That's a pallet. You want a subclassification, or is that good enough? [TRANS, FS]
14	B	No, no, pallet's fine. [TR, TR]
15	A	Okay. [P]
16	B	A CTG. [TQ/TI]
17	A	CTG is more than 1.
18	B	Oh, Okay. 105 SMK. [P, P, TI]
19	A	Is that a CTG 105 SMK?
20	B	It is indeed.
21	A	Okay. 2 pages of CTGs. CTG 105 ... SMK? [P, SELF, TQ]
22	B	Yeah [TR]
23	A	SMK, that's a pyrotechnic. [TRANS]
24	B	Okay, and 105 WP. [PC, TQ]
25	A	105 WP. [E]
26	B	A CTG 105 WP. [ADD]
27	A	Let's see. An APE or HE? Would it help if I read this to you? [P, TQ]
28	B	Alright, makes sense in certain ways. [TR]
29	A	There's the WP, I'm sorry. It's in pyrotechnic also. [P]]
30	B	Okay. [P]
31	A	I can tell you what's in ammunition if that would help. We've got a CTG 106 APE.
32	B	Okay. [P]
33	A	CTG 105 HE. [TI]
34	B	1 or 2? [TQ]
35	A	Both. Both 1 and 2. Then also in ammunition I have a CTG 40HE and a 60HE. [TR, ADD, C, TRANS]
36	B	A 60HE. [E]
37	A	Yeah. It seems to be a 60HE. [P]

M(essage) 1 is S(entence), 7 words long. The rest is ADD(ed information). B's first message contains a P(hatic) and S of 6. M3 is a P. M4 is either a T(erse) Q(uestion) or T(erse) I(nformation). Next we have an E(cho), followed by a P and a S of 9. M6 is C(onnector) and TQ. M7 is either INT(errupted) or TRUN(cated). M8 contains a F(alse) S(tart) and a S of 6. M9 is a P and an S of 3. M10 is a S of 3, however on

semantic grounds it could be considered an echo. The rule was adopted that a sentence echo was considered a sentence. M11 is either P or T(erse) R(eply), more likely the former, but the analysis in general would not be greatly affected by either choice. M12 starts with one or two Ps, more likely two, has a C and TQ. M13 is a TRANS(posed sentence), followed by FS and S of 3. Next we have either a S of 8, or two Ss, one of 3 and one of 4 and a C. Such sentences are fortunately infrequent. The general tendency was to separate such sequences unless semantic ties were strong. Again the influence on the overall analysis would not be great. M14 contains two TRs or a TR and a P, and a S of 3. Next line is a P. M16 is again either TQ or TI. Next is a S of 5. Next line is two Ps and TI, next two are Ss of 6 and 3 respectively, although the latter could be considered a phatic. M21 is a P followed by SELF(talking to oneself) and a TQ. Next is a TR. Next line is a TRANS. M24 is a P, C and TQ, next line is E. M26 is ADD. M27 is a P, a TQ, followed by a S of 9. Next a TR and a S of 5. It is problematic whether this should be a S. There are a number of possibilities. It could be P, could be ADD. Not many such decisions fortunately had to be made. The presence of the verb and the idiomatic character weighed toward sentencehood in this case. M29 is a S of 3, a P, a S of 4. M30 is P. M31 is a S of 11, followed by a S of 6. The former is typical of complex sentences with strong semantic ties. Next M is P, next TI, next TQ, next TR, ADD, C and TRANS of 13. Next is E, and the last one a P followed by a S of 7.

The working definitions for fragments and phatics are:

TQ (Terse Question): An elliptical question usually containing no VP, but often having a NP, e.g., "Why?", "How about pyrotechnics?" ("How about NP?" is quite common), "Which ones?".
TR (Terse Reply): An elliptical reply, also often just a NP, e.g., "No.", "Probably meters.", "50 and 7.62.".
TI (Terse Information): A rather elusive category, neither question, reply nor command, an elliptical statement but one often requiring an action. Examples can be appreciated in context only (Figure 1). It brings to mind Austin's How to Do Things with Words.[9]
E (Echo): An exact or partial repetition of usually the other speaker's string. Often an NP, but it may be an elliptical structure of various forms. A distinction was made at an earlier time between echo, self-echo, and echo-question but was abandoned. Only fragmentary echos (rather than whole sentences, which were far less common) were included.
ADD (Added Information): An elliptical structure, often NP, used to clarify or complete a previous utterance, often one's own, e.g., "It doesn't say anything here about weight, or breaking things down. Except for the crushables.", "It's smaller. 36"X20"X17".". Spelling out words was included here.
TRUN (Truncated): An incomplete utterance, voluntarily abandoned.
INT (Interrupted): One involuntarily abandoned. These two are often hard to distinguish, but truncation is clear if the speaker abandons his utterance, e.g., "Uh, some of these are ... I don't know what category they will go in.", and interruption is clear when one speaker jumps over the other's utterance which shows signs of intent at continuation, e.g., "A: Maybe we should work on some of the bigger things. B: Yeah, I think that A: Let's try some of the bigger decks. ...".
FS (False Start): These are also abandoned utterances, but immediately followed by usually syntactically and semantically related ones, e.g., "They may, they may be identical classes.", "Well, the height, the next largest height I've got is 34.".
COMP (Completion): Completion of the other speaker's utterance, distinguished from interruption by the cooperative nature of the utterance, e.g., "A: I've got a lot of...I've got B: 2 pages. A: Yeah.".
CORR (Correction): This may be done by either speaker. If done by the same speaker it is related to false start, but semantic considerations suggest a correction, e.g., "Those are 30, uh, 48 length by 40 width by 14 height.".
SELF (Talking to Oneself): Fragments, sometimes mutterings, even to the point of undecipherability, not intended for the other person, but rather thinking aloud reminiscent of Piaget's "collective monologue",[10] e.g., "Ummm - 7 7 8 5 and 14 - 7 7 8 will certainly add up to 22 wouldn't it or I guess.".
P (Phatics): The largest subgroup of fragments whose name is borrowed from Malinowski's[11] term "phatic communion" with which he referred to those vocal utterances that serve to establish social relations rather than the direct purpose of communication. This term has been broadened to include all fragments which help keep the channel of communication open, such as "Well", "Wait", but even "You turkey". Two subcategories of phatics are:
C (Dialogue connectors): Words such as "Then", "And", "Because" (at the beginning of a message or utterance).
T (Tag questions): e.g., "They're all under 60, aren't they?"

In the discussion above, the words "speaker" and "utterance" were used; but since most of these fragments are found also in the terminal-to-terminal mode and some also in the computational mode, they apply also to typed interactions.

3. Statistical Analysis of the Three Modes

The analysis here is based on the 1979/80 experiments only since they all involve the same shiploading task. The results were scored in each case by at least two persons, and the computational mode protocols by five. There are 8

face-to-face, 4 terminal-to-terminal and 21 human-to-computer protocols, involving 44 subjects. The time, was one hour each for the first two modes, and an average of one and one half hours for the third. Since there were twice as many F-F protocols as T-T and almost twice as many H-C as the first two combined, statistical totals are not very important. They are given here however to yield strength to the final processed comparisons.

The analysis of computational protocols clearly necessitated some different methodologies, and some data is simply not comparable (e.g., load sequences, since they were absent in F-F and T-T). The category "message" was split into "parsed message" and "parsed and nonparsed message," the first comprised of parsed inputs and the second of all inputs. The fragments also consisted of parsed ones: terse question, terse reply and definitions, and nonparsed ones: false starts and phatics. The terms "message" and "fragment" for the values in H-C refer to parsed messages and parsed fragments. Unless indicated otherwise, "fragments" in general do not include phatics, connectors and tags. Load sequences were completely left out of analysis, and obviously no computer answers were analyzed.

TABLE 3

	F-F		T-T		H-C	
Sentence length	6.8		6.1		7.8	
	(5.7-7.8)		(5.5-6.7)		(5.5-10.2)	
Message length	9.5		10.3		7	
	(6.4-12.4)		(7.8-12.7)		(total:7.8)	
Fragment length	2.7		2.8		2.8	
% of words in sentences	68.8		72.8		89.3	
% of words in fragments	17.2		21.1		10.7	
Sentences/message	.96		1.22		.81	
Fragments/message	.59		.74		.19	
Phatics/message	1.1		.59		.04	
	Total	Avg.	Total	Avg.	Total	Avg.
Messages	5574	697	310	78	1093	52
parsed and nonparsed					1615	77
Sentences	5302	663	385	77	882	42
Fragments	3253	402	230	58	211	10
Phatics (including connectors and tags)	4842	605	148	37	46	2
	Total		Total		Total	
Words in messages	49800		3285		8525	
Words in sentences	34266		2393		6880	
Words in fragments	8584		694		823	

The statistics show some expected marked differences as to the number of words, messages, sentences, fragments and phatics. The face-to-face mode is not surprisingly much more verbose, and shows a much higher ratio of phatics. What is however far more interesting is that several statistics are close to each other: those for sentence length, message length, fragment length (excluding definitions in H-C, since they are absent in the other two), percentage of words in sentences, especially for F-F and T-T, percentage of words in fragments, again especially for F-F and T-T. The latter two are of interest since in the H-C mode the percentage of words in sentences is higher and in fragments is lower, even though the system allows use of fragments. As for sentence length, Chafe[10] cites the "idea unit" in spoken as having a mean length of about 6 words. These numbers bring to mind George Miller's[12] "magical number 7". Also noticeable is a striking closeness between average of messages in T-T and parsed and nonparsed inputs in H-C. The ratio of sentence/message are close for the 3 modes, and the ratios of fragment/message are close for F-F and T-T. Nor surprisingly, the ratio of phatic/message are different, being particularly low for H-C.

Fragments are of particular interest and therefore are analysed in further detail. Fragments are considered separate from phatics. Nonparsed fragments in H-C are included in this analysis. TRUN and INT are collapsed into TRUN. As Table 4 shows TR is the predominant fragment in all three modes. (H-C mode characteristics are discussed in Section IV.) The next is ECHO for F-F, TI for T-T and TQ for H-C, and TQ is rather high in all three modes. These may seem to have little in common, but they are all typicaly NPs. The percentages for FS are close in all three modes, particularly so in F-F and H-C. The absence of some categories in some modes is equally interesting, even though totally understandable in some cases. The low presence of CORR in F-F and its absence in T-T is suprising, but may be partly due to some overlap of this category with FS. The absence of SELF and TAG in T-T and H-C is understandable, as is the absence of DEF(definitions) in F-F and T-T. It should be noted that in T-T the category did occur in a way. The subjects used a good deal of abbreviation in spelling (a common type of DEF is abbreviation) and also conventions, which every pair invented for end of message signal. ECHO and COMP in H-C would be rather silly -- who would echo or complete the computer? But the absence of ADD, CORR, TI and CON is due to the restraints of the grammar. Their role and desirability in H-C is further discussed in Section V.

TABLE 4

	Total	%	Length
Face-to-Face:			
Echo	532	16.4	2.7
Added Information	425	13.1	2.7
Correction	56	1.7	2.7
Completion	95	2.9	2.7
Talking to oneself	114	3.5	2.7
Terse Reply	571	17.6	2.7

Terse Question	411	12.6	2.7
Terse Information	297	9.1	2.7
False Start	413	12.7	2.7
Truncated	339	10.4	2.7
Definition			
Phatic	4842		1.4
Dialogue Connectors	1936		1.4
Tag Questions	31		1.4
Terminal-to-Terminal:			
Echo	10	4.3	2.8
Added Information	41	17.8	2.8
Correction			
Completion	2	.9	2.8
Talking to Oneself			
Terse Reply	67	29.1	2.8
Terse Question	31	13.4	2.8
Terse Information	48	20.9	2.8
False Start	23	10.0	2.8
Truncated	9	3.9	2.8
Definition			
Phatic	148		1.3
Dialogue Connectors	34		1.3
Tag Questions			
Human-to-Computer:			
Echo			
Added Information			
Correction			
Completion			
Talking to Oneself			
Terse Reply	91	37.8	1.0
Terse Question	67	27.8	4.6
Terse Information			
False Start	30	12.4	2.3
Truncated			
Definition	53	22.0	6.0
Phatic	46		2.3
Dialogue Connectors			
Tag Questions			

Phatics deserve a separate detailed discussion on account of their varied semantic functions but it is beyond the bounds of this paper. By far the most common phatic is Okay. It is interesting that speakers do not seem to be aware of this. When I asked my class in psycholinguistics (over 15 students) which phatic they thought most frequent, a variety of answers was given, but none came up with Okay. Table 5 shows the percentages of the top 5 phatics. In H-C several phatics occurred, but only 3 "Okay"s and one "Oh well" of the tope five. They are illustrated below and discussed in Sections IV and V. Table 5 also gives percentages for the top five dialogue connectors. There are none in H-C.

TABLE 5
Most Frequent Phatics

	F-F	T-T	H-C
Phatics			
Okay	27	25	7
Well	9		1
Uh	8		
Oh	7		1
Yeah	7		
Connectors			
And	33	28	
So	28	25	
But	10		
Then	7		
Now	5		

Some interesting phatics:
From F-F:
goddammit, bleah, oops, forget it, you're kidding, fool, yuk, you nitwit, what a pity, just a sec.
From T-T:
bleep, more to come, ook, ook to you, congrtltns, cmt => grt idea, stand by, you turkey ("look" occurred in 3 protocols, which is quite interesting considering the mode).
From H-C:
yes, I know how you feel, no, are you a computer?, of course, ?, foo to you, what is your problem?, there must be a better way, bla...bla, why don't you understand my question? help, where are we machine?, you lie, good, thank you.

IV. The Human-to-Computer Mode: Special Characteristics

1. Performance of the System

The system performance was such that meaningful work could be accomplished by largely uninitiated subjects with a bare minimum of assistance. Response to inputs which were not understood was extremely fast, the incidence of bugs was low (out of 1615 messages, 12 hit bugs) and recovery from them was excellent. Response times were quite adequate, especially since many requests involved quite a bit of computation. The subjects never showed impatience or boredom, but apparently used the latency time (from input to response) to formulate the next request.

2. The Influence of the Specific Task

The special task at hand and the special character of a problem solving situation both have an influence on the performance of the subjects. The "prompt sequence" for loading the ship provided in the language was used by all subjects even though they could have accomplished the same thing by natural dialogue (the magical number "7" shows up again here in the average of 7.6 loading sequences per protocol). The percentage of items loaded is lower than in the F-F but this is due to the considerably longer initial orienatation period in H-C (from 1/2 to 1 hour), after which the rate of loading increases. About 50% of items were loaded in F-F in one hour, so the task is completable in about two hours. About 20% of the items were loaded in H-C, but considering that the rate of loading increased in the last half hour of the sessions, the task was also doable in about 2 hours. The solution of the problem was not however of interest in these experiments. The influence of the problem solving situation was

very evident, particularly on syntax. The question (request) -- response interchanges are dominant in all modes. Rather short sentences used are also attributable to this. Fragments are useful for increasing the flow of information. Phatics facilitate interaction.

3. Syntax

The types of sentences used is of particular interest here, so detailed analysis was made with respect to sentence structure and type. The results are summarized in Table 6.

TABLE 6
Sentence Types

	Total	%
All sentences	882	
Simple sentences, e.g., "List the decks of the Alamo."	651	73.8
Sentences with pronouns, e.g., "What is its length?", "What is in its pyrotechnic locker?"	30	3.4
Sentences with quantifier(s), e.g., "List the class of each cargo."	71	8.0
Sentences with conjunctions, e.g., "What is the maximum stow height and bale cube of the pyrotechnic locker of the AL?"	88	10.0
Sentences with quantifier and conjunction(s), e.g., "List hatch width and hatch length of each deck of the Alamo."	23	2.6
Sentences with relative clause, e.g., "List the ships that have water."	6	.7
Sentences with relative clause (or related construction) and comparator, e.g., "List the ships with beam less than 1000."	6	.7
Sentences with quantifier and relative clause, e.g., "List height of each content whose class is class IV."	2	.23
Sentences with quantifier, conjunction and relative clause, e.g., "List length, width and height of each content whose class is ammunition."	2	.23
Sentences with quantifiers and comparator, e.g., "How many ships have a beam greater than 1000?"	3	.34

The dominance of simple sentences is striking. The reason is certainly not the lack of availability of complex sentences. I think that several reasons account for this. The problem solving situation influences the subjects to work in a simple manner, often employing what I have termed success strategy, i.e., repetition of the same type of requests. Another reason is definitions. Once the subject has introduced a definition whose right hand side is often complex, involving conjunctions, relative clauses, even quantifiers, they are used in subsequent requests, which are therefore short and simple. Another reason may be simply the computer. As Robinson[13] and Grosz[14] noted, subjects tend to be more formal in conversation with the computer.

Sentences were also analysed as to their type, since it was noticed that a great number of them were of the WH-type and contained be-verbs, e.g., "What are ships?". The results confirmed the observation: 75% were WH-type questions. Only 1% were Yes-No type questions, e.g., "Is Alamo a ship?", "Is there a deck whose primary use is ammunition and whose length is 396?". Commands, most commonly starting with "List", accounted for 19% of sentences, and a special category of statements, data addition, for the remaining 5%. These results are very interesting but I hesitate to offer an explanation. In the analysis of two F-F protocols consisting of 15500 words it was found that a be-verb occurred once every two sentences. Since be-verbs are so common also in F-F, this may either be a general feature of English or of the type of conversations in such problem solving tasks.

Concerning the occurrence of other verbs, few sentences contained HAVE-verbs. No other verbs were part of the version of the grammar available to the subjects. Verbs could have been introduced by definition, but nobody did so. Possessives and sentences with "there" were observed, but surprisingly few in view of the availability of these structures in the grammar. The use of the article "the" was erratic. The investigation of the F-F sample also showed few relative pronouns; "that" was the most common -- one in every 19 sentences. Conjunctions were fairly frequent -- one in every 8 sentences, "and" being the dominant one; likewise quantifiers -- one in every 10 sentences. This coincides well with the sentence analysis for H-C where sentences with conjunctions or quantifiers are the highest in percentage among the complex ones.

On the whole, one is forced to conclude that monotony of structure is the rule rather than the exception in H-C.

4. Definitions, Fragments and Phatics

The REL System allows the user to avail himself of a great variety of definitions[6] which, however, is not too well reflected in the protocols, due to the subjects' lack of familiarity with the system. One subject whom I observed as having familiarized himself with the system made extensive use of definitions. It should be added that, beyond those which were actually used, 30 more definitions were attempted but contained errors. Some definitions had been built in by the language designer, notably "remaining area" and "adjusted remaining area." These were frequently employed.

I have made a rough categorization of the definitions according to their complexity. Abbreviations are the simplest, e.g., "def:DKS:decks of the USS Alamo". But even abbreviations can be sophisticated and therefore more useful like the

following one with a quantifier: "def:ED:each deck of the Alamo." Abbreviations accounted for 34% of the total of 53 definitions. Synonyms were more complex: "def:INF1:aft width and forward width and minimum clearance," "def:INF2:INF1 and square foot capacity," "def:"well deck" info:INF2 of the "well deck" of the Alamo." Synonyms accounted for half (51%) of the definitions. Of the remainder, 9% involved arithmetical operations, e.g., "def:size:(length*width)/144", "def:F("8","9"):"8"*"8"+"9"*"9". A few definitions had to do with adding new data.

Other than definitions, fragments were of two types: parsed, which were Terse Question and Terse Reply, and nonparsed, which were False Starts and Phatics. TQs were noun phrases which are parsed into sentences if followed by a question mark, e.g., "Class of culvert?", "12*(SQ of MEZ)/(450/12)?" There are 67 of those. TRs were single words or numbers arrising from the particular feature provided by the system to deal with long answers. It reads, e.g., "There are 203 lines in this answer. How many do you want? Respond with 'all', 'none' or a number." It was considered important to include them, since failure to respond resulted in an error message, and also to see to what extent that feature is useful; it is, since there were 91 TRs. No distinction was made between False Start and Truncated; in all cases, these 30 occurences were messages abandoned by the subject for reasons that are seldom identifiable. A typing error may have been noticed or a thought changed, e.g., H: "What are the decks and primary uss" C: "Input Error" H: "What are the primary uses of each deck of the Alamo?" What is surprising about fragments is the paucity of TQs. They are handled by the system very well and are certainly shorter to type. I think that the reasons again are lack of familiarity with the system and more formal style on the part of the subject. But it is also possible that such elliptical structures are somehow more difficult to use, which would confirm transformational theory, but poses an uncomfortable question as to the desirability (widely assumed) of ellipsis in computational interaction.

Phatics are very peculiar in these H-C protocols. What is striking is the anthropomorphisation of the computer. This may be due to the background of the subjects, Caltech. They clearly also serve the function of venting one's emotions, and that may be useful. They are illustrated in Section III and number 46.

5. Special Strategies, Learning, Persistence of Errors

A number of interesting strategies with respect to the use of language were observed. The discussion here is just illustrative, but the annotation of the protocols shows that they were quite frequently employed. They are pretty self explanatory. (a) Paraphrase: e.g., H: "What do the DKS usually hold?" C: "Input error, please re-enter request" H: "What are the primary uses of the DKS?". Similarly: "How long is the Anchorage?" "What is the length of the Anchorage?". (b) Success: this usually involves repetitious structure of a sequence of requests, e.g., "What relations are there?" "What ship classes are there?" "Describe the AL." "Describe the DKS." "Describe water." "Describe tank.". (c) Simplification of Sentence Structure: e.g., H: "What is the maximum stow height,bale cube,square foot capacity and top stack number of the cube of the PL?" C: "Input error..." H: "What is the maximum stow height and bale cube of the PL?" C: "40 72" H: "What is the square foot capacity and top stack number of the PL?" C: "36 0". This example illustrates also the strategy of suspecting, and therefore changing, syntax. The subject made mistakes in punctuation by not using spaces, but changed syntax instead. That was quite a common strategy. Another example follows. (d) Mistrust of Syntax: e.g., "What is known about the hatch of each deck of Alamo?" "What is known about hatches?" "What is the deck of each hatch?" "What is known about hatch?" Each of these questions resulted in an error message, because the word "hatch" was not in the vocabulary. But the subject persistently changes syntax and morphology rather than vocabulary. In the following example, the subject misspelled the word "height", but still changes syntax first: "The heigth of the Alamo/39?" "What is the height of the well deck of the Alamo/39?" (e) Grammar Testing: this was not very frequent, but interesting: "What are the primary uses of each deck of the Alamo?" "What is the primary use of each deck of the Alamo?" (f) Data Addition: one subject added a new class: "item:=class", and then made all items to be loaded members of this class: "Water is an item." "Conwire is an item."..., thus enabling himself to deal with all items: "List dimensions of items."

Learning is evident throughout the protocols. It ranges from a simple correction of a misspelling or punctuation error to changing the syntax, for instance to obtain more specific answers; e.g., "What is it's beam?" (repeatd three times) "What is its beam?", H: "What is the minimum clearance of each deck?" C: "There are 224 lines ..." H: "What is the minimum clearance of each deck of the Alamo?"

Peristence of errors was occasionally found, as in the next to last example, and the following dialogue -- which shows a variety of errors and strategies such as success, mistrusting syntax, breaking the input into shorter ones, adding a quantifier instead of correcting punctuation. This sequence also illustrates some of the errors analyzed in the next subsection (6).

FIGURE 2

H: "C,D of lub oil SAE50"
C: "Input err..."(wrong punctuation, lack of space)
H: "C,D of sandbag"
C: "Input err..."(same)
H: "C,D of sandbag?"
C: "Input err..."(lack of space)
H: "List C,D of sandbag"
C: "Input err..."(wrong punctuation, space)
H: "List C of sandbag, lub oil SAE50, fuel oil, gen set PU670G."
C: "Input err..."(lack of conjunction before last NP)
H: "C of sandbag and lub oil SAE50."
C: "Input err..."(wrong punctuation)
H: "C of sandbag."
C: "Input err..."(same)
H: "C of sandbag?"
C: "Class IV"(good answer)
H: "C of lub oil SAE50,fuel oil,gen set PU670G?"
C: "Input err..."(lack of spaces and conjunction)
H: "C of lub oil SAE50?"
C: "Class III"(good answer)
H: "C of gen set PU670G?"
C: "Class II"(good answer)
H: "List D of sandbag, lub oil SAE50,fuel oil, gen set PU670G."
C: "Input err..."(lack of spaces and conjunction)
H: "List D of sandbag,lub oil SAE50, fuel oil, gen set PU670G?"
C: "Input err..."(lack of space and conjunction)
H: "List D of each sandbag, lub oil SAE50, fuel oil, and gen set PU670G."
C: "..."(good answer)

6. Analysis of Errors

Errors are all of those inputs that are not parsed, resulting in error messages. Not included in this analysis however are phatics and false starts, already covered. The message: "Input error, please re-enter request" is issued in all cases except when the system requests a response and the user fails to give one that can be understood, in which case the message is: "Your response is not understandable. Try again.", and in the case of bugs, where it reads: "Your request could not be completed as entered. Please rephrase and try again."

The total of 446 errors were classified into eight categories: (a) Vocabulary Errors: arising from the lack of a word in the language, e.g., "big" in "Is the Mobile a big ship?", "feet" in "List the decks of each ship with square feet capacity less than 70.". This being by far the largest category, the importance of the semantic factor is clear. (b) Punctuation: involves sentence final marks, commas and spaces; they are well illustrated in Figure 2. (c) Syntax: the low incidence of these errors is surprising; formal style, repetitiousness of structure, expediency in problem solving may all be factors. Errors involving conjunctions or prepositions are typical. Some difficult to categorize, nonparsed inputs were also included here, such as: "What is known?". In some cases, there are vocabulary errors but the syntax could not have been handled either, typically: "On what decks of the Alamo may cargo be stowed?", "stow" and "may" being not known. This input was immediately paraphrased as "What is the primary use of each deck of the Alamo?" and handled correctly; so one may wonder what is involved in cases which could not be reasonably expected to be handled. (d) Spelling: the only interesting observation is that some subjects noticed these errors immediately, others not for a while. (e) Transmission: terminal and phone line errors. (f) Definition Format: all errors in framing definitions are included here, whether vocabulary, punctuation or format. (g) Lack of Response: to "There are xx lines in this answer. How many do you want?" One subject tried 6 careful requests before catching on. (h) Bugs: the actual number of bugs encountered was very low. In a very few cases they resulted in termination of the session.

TABLE 7

	Total	%		Total	%
Vocabulary	161	36.1	Definition format	30	6.7
Punctuation	72	16.1			
Syntax	62	13.9	Lack of response	16	3.6
Spelling	61	13.6			
Transmission	32	7.2	Bug	12	2.7

In general, errors were far fewer and far different from what I expected. The high intellectual level of the subjects cannot account for that, since it was more than counterbalanced by lack of familiarity with the system and lack of knowledge of the task. What should be done about errors, and indeed what we are doing, is discussed in Section V.

V. Habitability and Naturalness of Human-Computer Interaction: Some Conclusions

The purpose of the experiments was to learn more about dialogues with the view to enhancing interaction with computers. What have we learned, and what are we doing? First, our guiding convictions have been confirmed: English, especially if augmented to suit specific tasks, is a natural and useful medium. The job of improving it is open-ended; English for the computer will never be all of English, since English is in reality not one language, but a variety of languages, among some of which all speakers choose freely, and many belong to specialists.

Our task is to build as good a system as our understanding permits, observe the results of its use in actual tasks, and then with increased understanding continue to improve. The REL System served well in the experiments; its rapid response time was well worth achieving

if for this purpose alone. But it is no longer a research tool. We are now building the POL (Problem Oriented Language) System.[15] What we have learned from the experiments is having a major influence on its design. Advances in parallel to our own are changing the human-computer relationship, and POL reflects these too.[7,16-19] Unlike REL, POL is programmed in a high-level language and thus more amenable for the research tasks that lie ahead.

System breadth and depth in Petrick's sense[8] and rapid response time remain our major concerns. Whatever improvements are introduced have to meet these requirements. Experiments leave no doubt as to their essentiality. Intelligent system response to the user, using his knowledge base, and support for building that knowledge base using the facile capabilities of English, are two major areas where changes are made.

Much is being done in the response to errors. REL was particularly weak in this area as Figure 2 on errors shows. Punctuation rules were too stringent, these can easily be relaxed and so designed as to almost entirely remove this source of error. For example, final punctuation can in almost all cases be added or corrected, and any ambiguities clarified gracefully. Even in REL "List ... ." and "List ... ?" are accepted, to the relief of users. Defaulted responses and responses that add additional information should be acccepted; for example, lack of response to "There are 203 lines..." caused 16 errors in the protocols, yet in the POL design it is handled by defaulting.

Identification of words not in the vocabulary and spelling correction did not exist in REL, resulting in a great deal of frustration. The two are related, and together accounted for 50 percent of errors. A problem here is the time inherent in spelling correction, however the new lexicon methods introduced in POL show promise of solving this problem. Syntactic and semantic means are used, as well as lexical, to identify intended usage, and echo is used to inform the user of the correction that is made; if the intent is not clearly identifiable, the user is informed, listing the troublesome words.

The users should be encouraged and guided to avail themselves of the wide range of definitional capabilities. This is a primary way for users to directly build knowledge into the system. Definition guides and help sequences are available in POL to this end. A major aspect of definitions is multiple defining of terms. To illustrate from one application of REL, the notion of "net sales" was defined in five ways; thus one could ask for "net sales of diodes","net sales of the Eastern Sales Region", "net sales of salesman Jones,"; the internal ambiguity was always clarified in context. However, the statistics from the experiments showing that of the 83 attempts at definition 30 were not sucessful point to needed improvements in making this capability available. I feel especially that the incorporation of verbs which are introduced by definitional paraphrase and which were used in other REL applications enhances naturalness, even though the experiments showed a preference for be-verbs.

The area of pronouns and ellipsis in general is, of course, very important. Pronouns worked to a certain extent in REL and they have been throughly revised for the POL System, profiting from the work of Grosz,[14] Sidner[20] and Robinson.[16] This area, however, will require much additional effort if we are to recognize the wide range of fragments - terse question, added information, and terse information. Some forms were handled by REL, e.g., "Dimensions of conwire?". However forms such as: "How about ..." and "Those of ..." need to be added. Added information might be handled in such a sequence as "Consider John, Joan, Betty and Bob. John and Bob are males. Joan and Betty are females. All are doctors." or "What is the longest tanker? Only Norwegian." Terse information and dialogue connectors may also be considered, for instance: "List the dimensions of vehicles.", and, following the answer, "And pallets."

Although I have only touched upon it briefly here, the prompt sequence in loading ships was an effective tool whose usefullness was strongly supported by the experiments. The setting up of such abbreviated means of communication by the user, as well as their use, will be supported in POL.

Finally, what about phatics? Should they be part of the computer's language? One is led by their wide use in face-to-face to include phatic messages from the computer, as is done in some of the other natural language systems. "Welcome," "Okay," "Thank you" are already in wide use. More of that nature would not hurt, within reason. Some inputs from the computer would undoubtedly be appreciated, such as: "Be patient, I'm working on it." if the computation is long or response delayed.

Is the recognition of users' phatics and response to them desirable? Fillmore[21] pointed out that politeness can be carried too far, as in the sequence: A: "You have lovely eyes." B: "Thank you." A: "You are welcome." Chafe[22] seems to be ready to see more human-like behavior on the part of the computer, even using variations in typing speed as a means of introducing a form of intonation and emphasis. We are currently investigating phatics, but while it could be interesting to observe users' reactions in this respect, naturalness may be more highly enhanced in other areas. And so, not knowing how to respond, swearing is likely to remain ignored by the forever imperfect computer.

References

1. Chapanis, Alphonse, "Interactive Human Communication," Scientific American, April 1957, pp. 39-42.
2. Martin, W. A. (Massachusetts Institute of Technology), personal communication.
3. Ervin-Tripp, Susan M., "Sociolinguistics," J. A. Fishman (ed.), Advances in the Sociology of Language, Mouton, The Hague, 1971, pp. 15-91.
4. Thompson, F. B. and Bozena H. Thompson, "Practical Natural Language Processing: The REL System as Prototype," Rubinoff, M. and M. C. Yovits (ed.), Advances in Computers, vol. 13, Academic Press, New York, 1975.
5. Thompson, Bozena H. and F. B. Thompson, "Rapidly Extendable Natural Language," Proc. of 1978 Nat. Conf. of ACM, pp.173-182.
6. Thompson, Bozena H., REL English for the User, California Institute of Technology, Pasadena, 1978.
7. Hendrix, G. G., "Future Prospects for Computational Linguistics," Proc. of 18th Annual Meeting of Asso. for Comp. Ling., 1980.
8. Petrick, S. R., "On Natural Language Based Computer Systems," IBM Journal of Research, vol. 20, no. 4, July 1976.
9. Austin, J. L., How to Do Things with Words, Harvard Univ. Press, Cambridge, 1962.
10. Piaget, J., The Language and Thought of the Child, Meridan, New York, 1955.
11. Malinowski, B., "The Problem of Meaning in Primitive Languages," Ogden, C. J. and I. A. Richards, The Meaning of Meaning, London, 1946.
12. Miller, G. A., "The Magical Number Seven Plus or Minus 2: Some Limits on Our Capacity for Processing Information," The Psychology of Communication, New York, 1967.
13. Robinson, Jane J., "Performance Grammars," Reddy, Raj (ed.), Speech Recognition: Invited Papers of 1974 IEEE Sym., Academic Press, New York, 1975, pp. 401-427.
14. Grosz, B. J., The Representation and Use of Focus in Dialogue Understanding, SRI International, Technical Note 151, 1977.
15. Thompson, Bozena H. and F. B. Thompson, "Introducing POL: A Problem Oriented Language System," Proc. of First Internat. Workshop on Nat. Lang. Comm. with Computers, Warsaw, Poland, Sept. 1980.
16. Robinson, Jane J., Diagram: A Grammar for Dialogue, SRI International, Technical Note 205, 1980.
17. Hendrix, G. G., E. D. Sacerdotti, D. Sagalowicz and J. Slocum, "Developing a Natural Language Interface to Complex Data," ACM Transactions on Database Systems, vol. 3, no.2, 1978.
18. Robinson, Ann E., D. E. Appelt, B. J. Grosz, G. G. Hendrix and J. J. Robinson, Interpreting Natural Language Utterances in Dialoges about Tasks, SRI International, Technical Note 210, 1980.
19. Hayes, Phil and Raj Reddy, An Anatomy of Graceful Interaction in Spoken and Written Man-Machine Communication, Computer Science Dept., Carnegie-Mellon Univ., 1979.
20. Sidner, C., Towards a Computational Theory of Definite Anaphor Comprehension in English Discourse, Artificial Intelligence Lab. Technical Report 537, Massachusetts Institute of Technology, 1979.
21. Fillmore, C. J., "A Grammarian Looks to Sociolinguistics," Shuy, R. (ed.) Report of 23rd Annual Round Table Meeting, Georgetown Univ. Press, 1973, pp. 273-287.
22. Chafe, W. L., "Should Computers Write Spoken Language?", Proc. of 18th Annual Meeting of Asso. for Comp. Ling., 1980.