Errors and usability of natural language in a multimodal system

J. Gregory Trafton and **Kenneth Wauchope** and **Janet Stroup**
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375-5337
{trafton,wauchope,stroup}@aic.nrl.navy.mil

Abstract

A pilot experiment was run that used direct manipulation, (typed) natural language, and a combination of direct manipulation and natural language as input modalities. Approximately 10% of the natural language utterances that users issued were not understood or could not be processed by the system. These utterances were categorized into "user errors" and "system errors," and then further analyzed and described. Also, general usability of the system was evaluated by an exit questionnaire. Finally, suggestions for future systems were discussed.

Introduction

Multimodal systems are very difficult and expensive to build and have many potential problems: How good does the graphics or natural language or sound system have to be? How should the multiple systems interact? What kind of errors do users make using each kind of system? What kind of errors does the system make? Do users have a preferred method of interaction?

One of the problems with current multimodal systems is that designers rarely look at the number or type of errors that users make while interacting with the system. In theory, it is usually possible to determine what kinds of utterances or interactions a system can not deal with, but users rarely make some kinds of errors and often make other types of errors. By examining the types of errors that people make when interacting with the system, it should be possible to reduce those types of errors, and create better, more user-friendly multimodal systems. If error types can be generated that are general enough to be consistent across different systems, designers can build systems that take these errors into account, and the resulting systems should be significantly better.

Unfortunately, there have been very few empirical studies that investigate these issues. The few studies that have been done typically use "wizard of Oz" strategies for the more difficult aspects of the interface (i.e., an experimenter responds for the parts of the system that are not yet implemented). In this study, participants were asked to perform a variety of tasks using a natural language system. This paper describes the errors that users made while interacting with the system as well as possible ways to prevent these types of errors from occurring in the future.

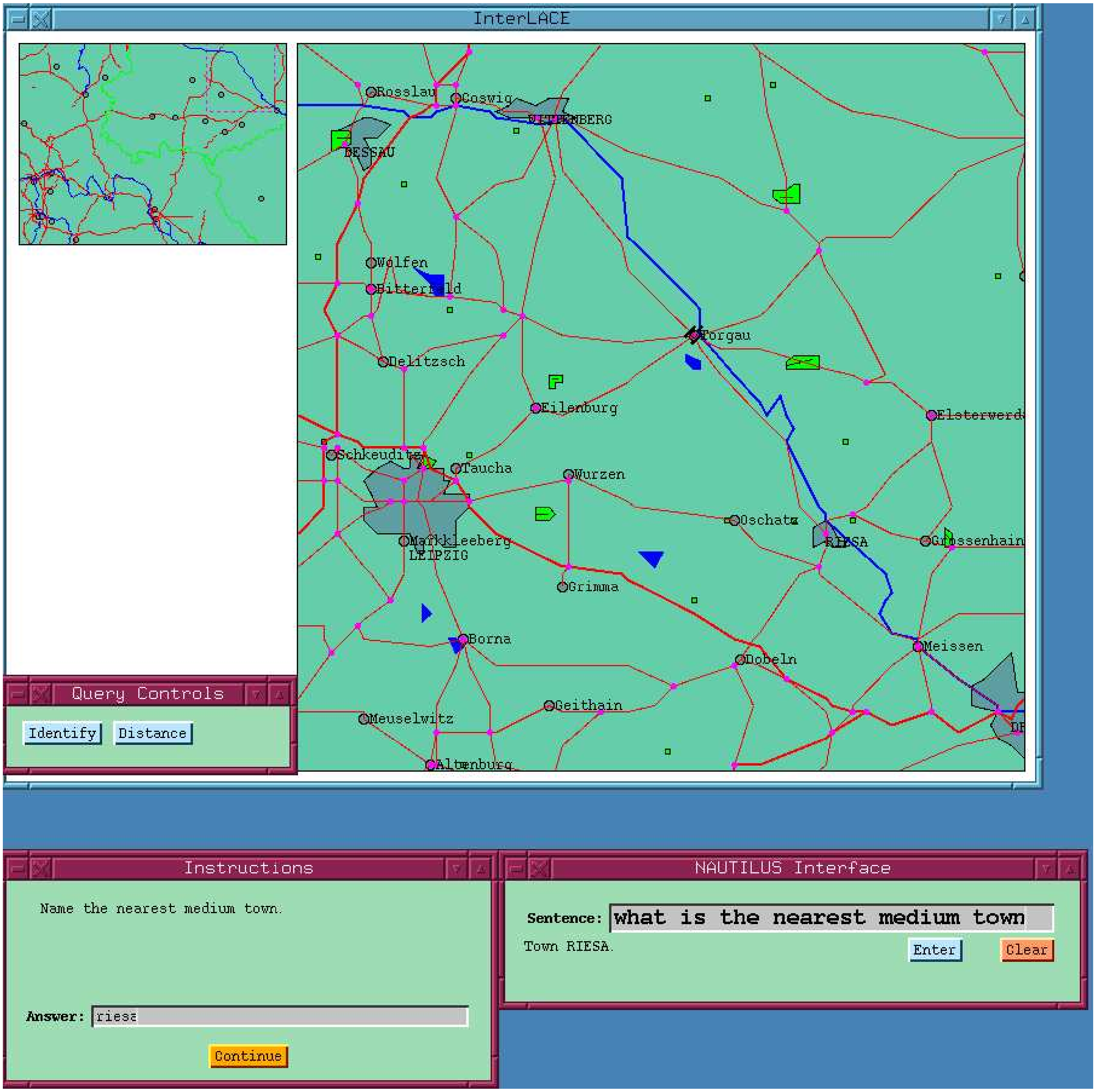
Method

We manipulated whether subjects could use (typed in) natural language, direct manipulation, or a combination of natural language and direct manipulation for system input to a cartographic system called InterLACE.

Participants were 24 volunteers from NRL. 9 participants were in the natural language (NL) condition, 8 participants were in the Graphical User Interface (GUI) condition, and 7 participants were in the combined (combined) condition.

InterLACE is a fully pannable, zoomable, mouse-sensitive graphical map display of southern Germany which has been interfaced to our natural language processor NAUTILUS to provide natural language capability. Figure 1 shows a screen snapshot of InterLACE in the combined condition. NAUTILUS (Wauchope, 1994, 1996) is a modular natural language processing system consisting of the PROTEUS syntactic analyzer from New York University (Grishman, 1986), the TINSEL semantic interpreter (Wauchope, 1990), FOCAL reference resolution component, and FUNTRAN quantified-expression builder. TINSEL, FOCAL, and FUNTRAN were developed in-house at NCARAI. NAUTILUS has been used as the natural language processor in several different applications prior to InterLACE, including Eucalyptus (an interface to a simulated air combat C2 station) (Wauchope, 1994), and InterVR (a speech controller for a 3D immersive battle simulation playback system) (Everett, Wauchope, & Pérez-Quiñones, 1994).

The PROTEUS language model contains 384 words, many of which are unused morphological variants automatically generated by the PROTEUS lexical macros.



Name the nearest medium town.

Answer: riesa

Sentence: what is the nearest medium town

Town RIESA.

Figure 1: A screen snapshot of InterLACE in the combined condition. The main window is the “work area;” the small window in the upper left shows the complete InterLACE “world;” the “Query Controls” are used with the mouse to get information and distance from objects on the map; the “NAUTILUS Interface” window is used to type arbitrary commands and queries; the “Instructions” window presents various tasks to the user.

The speech recognition model, by comparison, contains 160 words.

A simulated tank unit (positioned in Torgau in Figure 1) responds to typed in route instructions and mouse drags. In this experiment, subjects who used natural language to issue commands and queries typed sentences or parts of sentences into an input window (the "NAUTILUS Interface" window in Figure 1). Everything that could be done in one input-modality (i.e., natural language) could also be done in the other (i.e., direct manipulation). A more detailed description of InterLACE can be found in Wauchope (1996) and Trafton, Wauchope, Raymond, Deubner, Stroup, and Marsh (submitted).

Participants were presented with instructions like "Go to the intersection nearest town Fulda" or "Name the two roads that cross intersection 322." Users had to make the tank go to the location or query the system and enter the answer for all questions (see Figure 1 for a "query" instruction). There were 32 total instructions. All keystrokes and mouse-clicks were recorded and time-stamped. After the participants finished the task, an exit questionnaire was filled out to assess general motivation, impressions of the system, and how difficult it was to use the system. The exit questionnaire presented questions both open ended and in a lickert-scale (1—7 range) format.

Results and Discussion

Of the three conditions, only two (NL, combined) used natural language. Since this paper is concerned with the errors that were made by participants who used the natural language system, the GUI condition will not be discussed further (it should be noted that no "errors" were made by participants in the GUI condition).

Subjects in the NL and combined condition typed in an average of 50 sentences, and made an average of 4.8 "errors." An error in this experiment was defined as any sentence that NAUTILUS could not understand, parse, or execute. As Table 1 shows, the subjects in the combined condition used the natural language system less than subjects in the NL condition (because they also had access to the graphical system), $F(1, 14) = 26.6$, $MS_e = 88.3016$, $p < .0001$. There were no apparent qualitative differences between the types or proportion of errors that the two conditions made, so all future analyses will combine the two conditions.

Condition	NL steps	NL errors
NL	60.4	6.2
Combined	36	3

Table 1: Average number of natural language steps and errors made by the NL and combined conditions

In this experiment, almost 10% of the time, participants typed in something that InterLACE did not understand, could not parse, or could not execute. What were these errors? Were participants able to recover from them? Did they cause users to dislike the system? We examined these issues by categorizing the errors that participants made into "user errors" (errors that were caused by the user), "system errors" (deficits in coverage), and "miscellaneous errors" (described below).

User errors consisted of spelling or typo errors, ungrammatical errors, and telegraphic errors. System errors consisted of utterances that could have been understood and processed, but the participant used unexpected vocabulary and/or syntax. Miscellaneous errors included utterances that InterLACE could have understood, but were disabled for this particular experiment, and errors that were not procedurally defined in this system (i.e., attempting to go to an airstrip without a road leading to that airstrip). Table 2 shows the proportion of errors in each category.

Spelling / Typos

Since subjects had to type in their sentences, it is not surprising that some subjects made spelling errors or typos. Examples of incorrect sentences were:

`g0 to gersfeld` (notice the number zero)

`go to the clostest intersection to airstrip illesheim`

Fixing this type of error is a simple procedure (e.g., putting misspelled words in the lexicon), though it is very time consuming. It is possible to put very common misspellings in the lexicon (e.g., "artic") but that would never be able to deal with arbitrary typos like g0 and clostest, which need an interactive spell checker or something similar.

InterLACE also has the ability to accept verbal (spoken) input. If users spoke their commands, there would be no spelling errors, though there would perhaps be verbal speaker recognition problems (e.g., Damper & Wood, 1995). A future study has been planned to evaluate this issue.

Ungrammatical

Sometimes subjects issued utterances that were not grammatical, usually leaving out words. Examples of ungrammatical sentences were:

`go to fulda no intersection 321`

`go gersfeld`

For some of these types of errors, error correction might be possible (like inferring the deleted preposition in `go gersfeld`) but not others. For example, only by looking at the user's subsequent inputs and the goal she was attempting to accomplish was it possible to determine that `no intersection 321` meant "without going through intersection 321."

Error Type	Error Category	Percentage
User Errors	Spelling/typos	22%
	Ungramatical	9%
	Telegraphic	30%
System Errors	Outside present coverage	27%
Misc. Errors	Compound sentences	3%
	Not procedurally defined	9%

Table 2: The proportion of errors in different categories.

Telegraphic

One of the most interesting errors that users made was to speak "telegraphically," leaving out the determiner `the`. We suspect that users speak this way to InterLACE because they think that InterLACE "wants" to be spoken to in this manner (Don, Brennan, Laurel, & Shneiderman, 1992; Brennan & Ohaeri, 1994). Examples of incorrect sentences were:

`go to town nearest to intersection 381`
`what is nearest town`

Fixing this type of error is easy to do, though it does bring up issues of how "natural" the language is or should be. There are multiple examples of well defined telegraphic sublanguages (e.g., Fitzpatrick, Bachenko, & Hindle, 1986) that do determiner deletion in a uniform (and thus linguistically motivated) manner.

Compound sentences

Since the experiment was concerned with giving all conditions "equivalent" commands and abilities, compound sentences were disabled because a simple way of making this type of command available to the GUI condition was not available. InterLACE can process these sentences, but they were disabled for this particular experiment. Examples include:

`go east to road e70 and go to intersection 373`
`go to e4 and take it to kassel`

Clearly, this is not a true "error" though the system flagged it as such to the participants.

Not procedurally defined

Some aspects of the map were there for informational purposes and could not be traveled to. For example, airports and heliports were on the map, but the user could not actually take the tank to those places. For example:

`go to airstrip Illesheim`
`go to road e4 via road e70`

Errors of this type are rather difficult to fix, since it is not immediately obvious how to reinforce the fact that the tank can only travel on roads (which participants were told during the system demo).

Outside present coverage

These were the only true system "errors." Subjects sometimes used syntax or vocabulary that were not in the lexicon of InterLACE. For example,

`go nw` (system understands northwest, not "nw.")
`where is the nearest medium sized town`
`progress to bad-neustadt` (system does not understand "progress").

Correcting this type of error can be simple (i.e., making "nw" a synonym for "northwest" by adding it to the lexicon) or complex (i.e., changing the grammar and semantic interpreter for "medium sized town" — the system currently understands "medium town").

Exit Questionnaire

The exit questionnaire asked a number of questions concerning usability and ability to solve problems. The most relevant question for this paper was "As compared to other computer applications you are familiar with, how hard did you find it to interact with this application?" There were no significant differences between conditions on this question, $F(2, 21) = 0.53$, $MS_e = 3.4$, *n.s.*, though the means were very high: On a scale of 1 (difficult to interact with) to 7 (easy to interact with), means were 5.3, 4.9, and 5.9 for the NL, GUI, and Combined conditions, respectively.

Discussion

Overall, approximately 10% of the utterances that users made were not understood by the system. Of those, only 2.6% of the participants' utterances were actual "system" errors, or errors that were outside NAUTILUS's present coverage. The most common type of error made by users was speaking telegraphically, which is relatively easy to fix. Allowing users to speak instead of type, allowing users to speak telegraphically, and modifying the lexicon to allow common ungrammatical utterances would reduce over 50% of the errors identified in this study.

Did the number or type of errors that users experienced make the system less usable than a GUI system? Apparently not, as shown by the exit questionnaire. Users rated the natural language system very highly on an absolute scale, and gave the system high praise.

Reducing the number of utterances that are not understood by the natural language system to 3% or less is certainly a possibility and something most systems should strive for.

Acknowledgments

Greg Trafton and Ken Wauchope contributed equally to this research. The authors would like to thank Paula D. Raymond, Astrid Schmidt-Nielsen, and Elaine Marsh for comments on earlier drafts of this paper.

This research was supported in part by 6.1 funding from the Office of Naval Research to the Naval Research Laboratory and by problem number 55-5008-07 from the Office of Naval Research. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U. S. Navy.

References

Brennan, S. E., & Ohaeri, J. O. (1994). Effects of message style on users' attributions toward agents. In *Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems*, Vol. 2 of *SHORT PAPERS: HCI Research?*, (pp. 281–282).

Damper, R. I., & Wood, S. D. (1995). Speech versus keying in command and control applications. *International Journal of Human-Computer Studies*, *42*(3), 289–305.

Don, A., Brennan, S., Laurel, B., & Shneiderman, B. (1992). Anthropomorphism: From eliza to terminator 2. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*, Panel, (pp. 67–70).

Everett, S. S., Wauchope, K., & Pérez-Quiñones, M. A. (1994). A natural language interface for virtual reality systems. Technical Report NCARAI Internal Report AIC-94-046: Naval Research Laboratory, Washington, DC.

Fitzpatrick, E., Bachenko, J., & Hindle, D. (1986). The status of telegraphic sublanguages. In R. Grishman, & R. Kittredge (Eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, (pp. 39–51). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grishman, R. (1986). Proteus parser reference manual. Technical Report PROTEUS Project Memorandum #4: Department of Computer Science, Courant Institute of Mathematical Sciences, New York University.

Trafton, J. G., Wauchope, K., Raymond, P. D., Deubner, B., Stroup, J., & Marsh, E. (submitted). How natural is natural language for intelligent tutoring systems? In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*.

Wauchope, K. (1996). Multimodal interaction with a map-based simulation system. Technical Report AIC-96-027: Naval Research Laboratory, Washington, DC.

Wauchope, K. (1990). A tandem semantic interpreter for parse section. Technical Report 9288: Naval Research Laboratory, Washington, DC.

Wauchope, K. (1994). Eucalyptus: Integrating natural language input with a graphical user interface. Technical Report 5510-94-9711: Naval Research Laboratory, Washington, DC.