THE CONVERSING SEARCH ENGINE: SHOULD SEARCH ENGINES INTERACT WITH THE USER THROUGH CONVERSATION?

By

NELSON FRANCIS KIDD

A THESIS

Presented to the Department of Computer and Information Science
and the Honors College of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Bachelor of Science

August 1999

APPROVED: [redacted]

Stephen Fickas

An Abstract of the Thesis of

Nelson Francis Kidd for the degree of Bachelor of Science in the Department of Computer and Information Science to be taken August 1999

Title: THE CONVERSING SEARCH ENGINE: SHOULD SEARCH ENGINES INTERACT WITH THE USER THROUGH CONVERSATION?

APPROVED: [redacted]
Stephen Fickas

Previous research shows that anthropomorphic and natural language interfaces are feasible, efficient, and popular for information retrieval scenarios. However research on information retrieval interfaces capable of both natural language and emulation of human personality is not greatly explored. This study seeks to determine if a natural language interface, combined with a virtual personality, performs better than a traditional Boolean interface in the context of a search engine. The study also seeks to determine if people prefer the anthropomorphic interface to the traditional interface. People were asked to do information retrieval tasks using both interfaces and to evaluate the interfaces after using them. The results revealed that neither interface provided an advantage in terms of accuracy, but the traditional interface provided greater time efficiency. In agreement with previous studies, people showed more preference to anthropomorphism.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF GRAPHS

I. INTRODUCTION

1. Emergence of Agent Technology for the Casual User

The increased power of computer technologies has arrived at a level where computer-related tasks can be delegated to a computer. The software technologies that perform these functions on behalf of a person are usually termed *software agents*, or agents for short. Although some debate exists regarding the best definition for a software agent, such a definition simply gives context to this paper. Some definitions require advanced functionality in the software, but simpler definitions also exist. The definition in this paper is simple and focuses on the general intent of software agents – doing tasks for people.

The idea of an agent doing a task for a person is not new and is commonly found in science fiction and Hollywood's depictions of machines. Hollywood has given inspiration to many kinds of advanced agents, ranging anywhere from a talking computer in Star Trek to the robots in Star Wars. Although Hollywood may seem years ahead in its depictions of technology, we already see basic software agents in casual computer environments. Many software companies use *wizards*, a type of agent, to install software on various computer platforms. For those who are familiar with the Internet, the helpful tools known as *search engines* are also agents, whose job is to find documents on websites using keywords. Keywords are the words

provided by the user to the search engine. If a person wants to find something about *cats*, then the person would use a keyword of *cat* or *cats*.

2. Search Engines and their User Interfaces

An important component to the design of any software is the *user-interface (UI)*. The user-interface is the mechanism through which users interact with a software application. Most software applications use a user-driven graphical user interface (GUI). Users of GUI applications have windows, pull-down menus, dialog boxes, toolbars, buttons, and other graphical elements to control their interaction with the software. The *user-driven* aspect of an interface simply means that the user controls the program while it is running, often through a keyboard and mouse, while receiving feedback from the computer through a monitor. In user-driven environments most of the software's responses are results of the user's actions. An alternate form of interaction between user and software is *mixed-initiative*. Mixed-initiative environments allow software to take the initiative, thus making the user respond to what a computer does. Although mixed-initiative agents exist, all of the software in this paper concerns itself with user-driven software.

The predominant interface design for most software applications is based on a user-driven GUI model. Common examples of such applications are Internet web browsers like Netscape Navigator, as well as business applications like Microsoft Office. Internet search engines also make use user-driven GUIs. These search engines usually have a textbox where a

person types keywords, selects the domain[1] for the search, and submits the subject words for the search. The search engine then returns a list of documents that matched the keyword criteria. The screenshots shown in Figure 1 and Figure 2 are examples of traditional Boolean[2] search engine interfaces. Figure 3 shows a search engine that appears to use a traditional Boolean interface, but in reality uses a natural language interface. Instead of typing actual keywords, users can type actual questions in plain English.

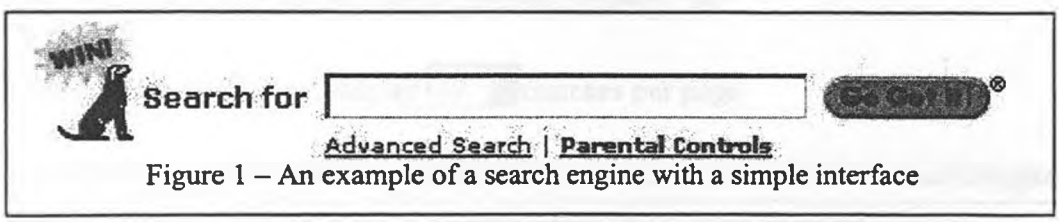


Figure 1 – An example of a search engine with a simple interface

[1] A search domain is simply a category of information that can be searched. For instance some search engines allow you to search through webpages, file databases, message forums, or more.

[2] Boolean algebra uses AND, NOT, and OR operators to show relationships between keywords. Hence the keywords *cat* and *dog* could be connected by the AND operator yielding the Boolean expression "*cat* AND *dog*". Such an expression would tell the search engine to look for things about cats and dogs.

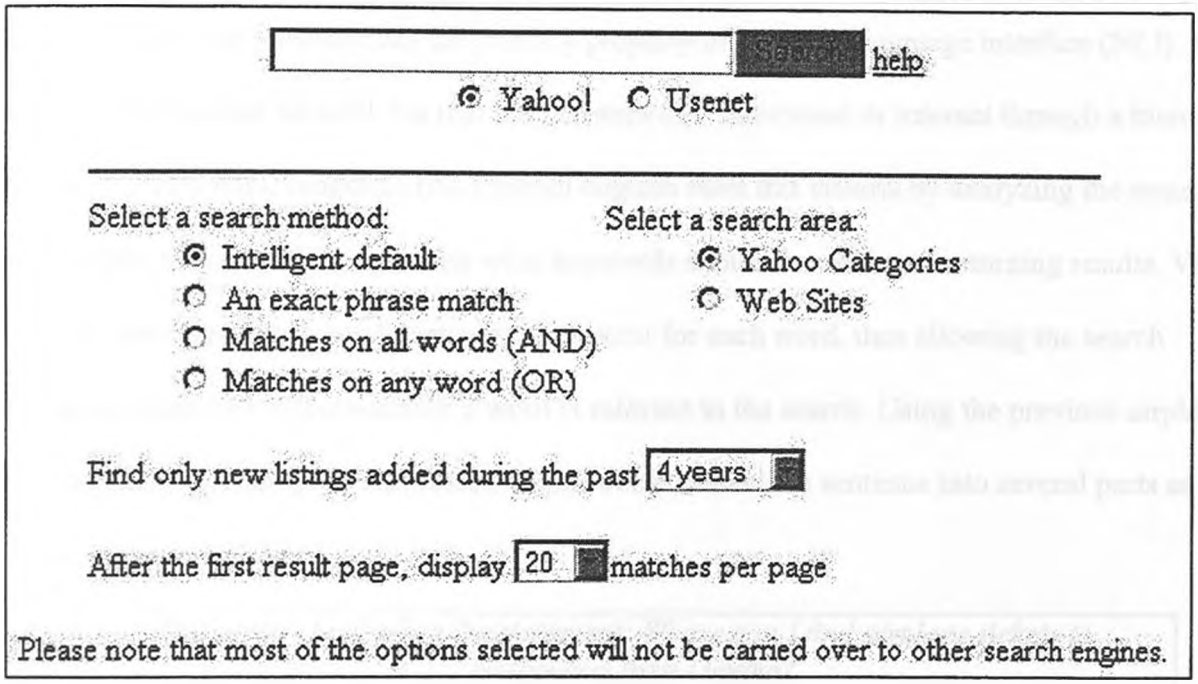


Figure 2 – An example of a search engine with a more complex interface

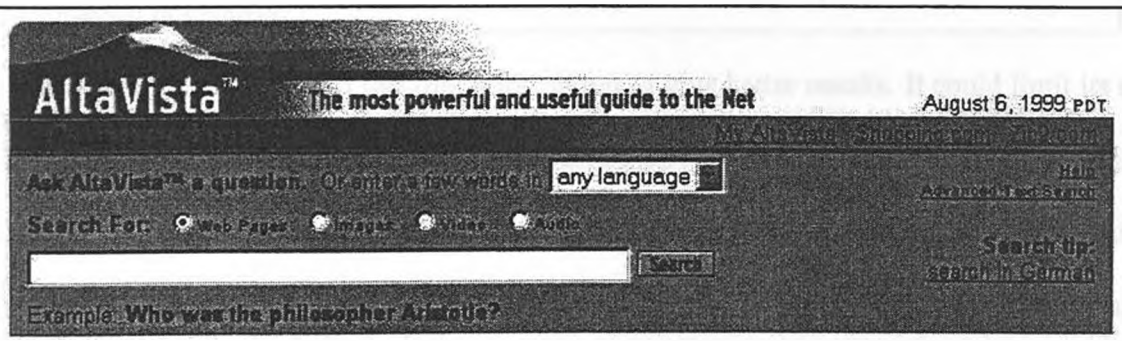


Figure 3 – An example of a search engine that uses natural language queries while maintaining the look and feel of a traditional interface.

Some search engines, like those provided by www.ask.com and www.altavista.com, employ limited natural language capabilities[3]. For example people could type the following request:

[3] Ask Jeeves, Inc. and AltaVista own www.ask.com and www.altavista.com respectively.

Where can I find airplane tickets to Hollywood from Omaha? Although the users still do this through a GUI, the interface has the primary property of a natural language interface (NLI). The primary requirement for a NLI is that the software can understand or interact through a human language. The natural language (NL) search engines meet this criteria by analyzing the sentence provided by the person, determining what keywords should be used, and returning results. Very complex search engines can determine the context for each word, thus allowing the search engine to better determine whether a word is relevant in the search. Using the previous airplane tickets example, a complex NL search engine could dissect the sentence into several parts as shown Figure 4 below.

Figure 4 – Analyzing the statement: *Where can I find airplane tickets to Hollywood from Omaha?*

Subject: I
Verb: find
Direct objects: airplane tickets
Word relationships: The ticket destination is *Hollywood* and the departure location is *Omaha*.

A complex search engine could use this information to give better results. It could limit its search only to things about airplane tickets, and then narrow down that search for flights departing from Omaha and arriving in Hollywood. On the other hand a simple natural language search engine might examine the statement to look for anything relating to airplane tickets, Hollywood, and Omaha.

One might ask, *where is all of this leading*? Search engines (and software in general) are approaching the realm of the talking computer. Natural language interfaces are starting to get attention, as computers are finally able to understand language like people. With many new technologies coming into development, new questions come to mind: *Are efficient and accurate natural language technologies possible? Are they worth the time and development? When are*

natural language interfaces useful? Should daily items like microwaves and televisions be controlled through natural language speech? Various people in the field have offered their thoughts and debated those questions. At least for the purposes of this article, I will present some of the notable successes for natural language interfaces involving information search and retrieval.

II. PIONEERED GROUND

1. Natural language Interfaces Work Well

Although less effort is directed towards the evaluation of natural language and agent technologies, some recent evaluative studies have been conducted showing that natural language is a viable method of doing information retrieval (IR) tasks. Capindale and Crawford observed casual users interacting with a real-life database through a natural language interface. The NLI had a restricted grammar appropriate for the database, and many of the user errors were related to the vocabulary restrictions of the NLI. Despite these problems users used feedback from the interface to avoid and recover from errors. Through their observations Capindale and Crawford made two important conclusions:

> *Natural language processing technology is developed enough to handle the limited domain of discourse associated with a database; it is simple enough to support casual users with a general knowledge of the database contents; and it is flexible enough to assist problem-solving behaviour. (Capindale, Crawford, 1990)*

> *Results show that natural language is an efficient and powerful means for expressing requests. This is especially true for users with a good knowledge of the database contents regardless of training or previous experience with computers. Users generally have a positive attitude towards natural language.*
> (Capindale, Crawford, 1990)

Building upon the conclusions of successful NLIs at simpler levels, Howard Turtle compared natural language systems and commercial Boolean search methods. His experiment examined the results of leading natural language search systems against other commercial search systems that used Boolean query languages. The results showed that "natural language searches consistently produced better rankings." (Turtle, 1996) He also added that the Boolean queries used in the experiment were created by "expert searchers who are intimately familiar with the collections[4] involved, which suggests that the improvements over an "average" or novice user would be even more pronounced." (Turtle, 1996)

Another claim supporting the viability of natural language is that "natural language processing can now be done on a fairly large scale and that its speed and robustness has improved to the point where it can be applied to real IR[5] problems." (Strzalkowski, Perez-Carballo, 1996) A survey published in 1997 concluded that the development of successful and effective natural language information systems in businesses required careful attention to certain stages during development. The researchers noted those stages:

1. *Systematic analysis of the company's requirements.*
2. *Effective integration of the natural language technology with the target database ensuring current applications are not adversely affected.*
3. *Introduction to new users of the system. This resulted in realistic user expectations and enabled effective use of the natural language software.* (Sidhu, Hinde, 1997)

This survey admitted that natural language interfaces are not uniformly successful, but a general conclusion is that effective and accurate natural language interfaces are possible and viable solutions provided careful planning is involved.

---

[4] *A collection refers* to the contents of the database.

[5] IR is short for Information Retrieval.

2. Do NLIs that Support Anaphora and Ellipsis Work Well too?

One of the characteristics of most natural language search engines in the experiments is that they do not employ *anaphora*[6] or *ellipsis*[7] technology. Both of these technologies have been explored, and the conclusions state that NLIs should be functional and flexible but limited in their language capabilities. One previous set of experiments compared an ordinary conversational format with a simulated automated dialogue system. In both dialogue models, the computer's dialogue was actually provided by a human. The conclusions support the hypothesis that "unrestricted human-human conversation is not the most appropriate model for the design of human-computer dialogue interfaces" (Ringle, Halstead-Nussloch, 1989). This claim implied that natural language interfaces should have formalized limitations, and the interaction between human and computer should be easy for the users to follow.

Following the need for formalized and restricted interfaces led to two important studies about NLIs in 1993. The first study examined whether or not natural language interfaces need a sublanguage to help formalize restrictions. A sublanguage is a language "reduced in complexity, but nevertheless providing the flexibility of natural language." (Burton, Steward, 1993) Such a sublanguage would allow people to use plain language without running into the problems of ambiguities of spoken language. The study summarized the findings in the following statement:

---

[6] A phrase that refers to something mentioned earlier. For example if somebody asked a search agent: *Where can I find a flight to Hollywood?*, the search agent might respond with: *From where would you like to depart?* The user could say: *I would like the flight to depart from Omaha.* An anaphora in this example is that *the flight* in the last sentence is the same *flight* in the first sentence.

[7] Ellipsis involves the use of sentences that appear to be incorrectly formed. Usually the parts that are missing can be extracted from the previous sentence. For example: *Where are the bananas? The peaches?*

> *In other words, a simple, restricted, clearly delimited, and perspicuous sublanguage of English, supported by meaningful feedback and error messages, will be sufficient for natural language database query.* (Burton, Steward, 1993)

This conclusion is important as it adds support to the previous claim that a natural language interface need not support the full range of human language to be successful, provided that the user is given useful messages from the software. The other study used a NLI and three additional variants of the chosen NLI. One variant allowed an extralinguistic[8] enhancement that allowed users to recall a previous statement and edit the query. The second variant had elliptic technology built into it. The third variant had both the recall-and-edit and elliptic technology. The study then compared the four different versions of the natural language query system in information retrieval scenarios.

> *The results show that enhanced linguistic capabilities can indeed improve usability under certain circumstances, but that extralinguistic enhancements can be just as effective. The results also show that usability can actually degrade as both the linguistic and the extralinguistic capabilities of an interface are improved.* (Burton, Steward, 1993)

These previous experiments are important because they support the theory that it is not necessary to make natural language systems handle all of the complexities of daily natural language.

3. Virtual Personalities Work Well

So far I have only discussed agents with respect to natural language interfaces. Another area of interest to agent technology is the issue of virtual personalities[9]. Hollywood has inspired us with robots and computers with personality, and computer scientists have also looked into

[8] The extralinguistic enhancement was provided by a GUI component for a recall-and-edit feature.

making the virtual personality into a reality. Erickson examined design issues of an agent's metaphor (human characteristics) and its ability to adapt to a user. He concluded that the key to designing agents is to "minimize the impact of errors and to enable people to step in and set things right as easily and naturally as possible" (Erickson, 1997). With regards to creating effective agent metaphors, Reilly states:

> *By focusing on a small set of important elements of a personality, the problem of creating personality-rich characters becomes more concrete. By following a methodology that suggests creating behaviors that are tailored to the character and environment and that use minimal amounts of representation of other characters, I have been able to create "working" behaviors...* (Reilly, 1997)

One study concluded that animated agents with lifelike characters can make a positive impact on the learning process. (Lester, Converse, Kahler, Barlow, Stone, and Bhoga, 1997) With such conclusions being applied to agent technology, the next step in creating the "talking" computer seems to involve bestowing a personality to a natural language interface.

4. Do Virtual Personalities Work Well with Natural Language Query Systems?

Previous work shows that natural language interfaces and virtual personalities are both possible, but they have yet to be evaluated together in the context of an information retrieval system. Natural language systems function well, but their performance becomes questionable when language enhancements, like ellipsis and anaphora, are allowed. Virtual personalities have a positive impact in arena of learning and tutoring, but they have yet to be evaluated with searching software. Adding a virtual personality to a natural language interface creates an extended natural language interface. I define such an interface as a *conversational interface.*

---

[9] Virtual personalities are human-like personalities applied to computers and software *Agent metaphor* is sometimes

It is my intent to evaluate the efficacy of a *conversational interface* in areas of performance and user satisfaction through information retrieval tasks. The specific questions I intend to examine are

- Is a conversational interface faster than a traditional interface?
- Is a conversational interface more accurate than a traditional interface?
- Do people prefer a conversational interface to a traditional interface?

While one may speculate that conversational interface would be better than a traditional interface according to Howard Turtle's research, Turtle's research looks at very advanced information retrieval systems. The Boolean search engines in his study used advanced Boolean grammars and the natural language interfaces can decipher the context of words. The underlying search engine in this study, inspired largely by the Internet search engines, is much simpler than the industry-level software in Turtle's research.

The conversational interface in this study (ELISE) is a simple virtual personality that utilizes a few anaphoric and elliptic capabilities within a natural language search engine interface. Although it has previously been shown that anaphora and ellipsis are not required (Burton, Steward, 1993), a successful virtual personality needs some capability to understand anaphora and ellipsis for a conversation. Despite this enhancement, ELISE is still a primitive search engine. Unlike some of the advanced natural language systems in Turtle's study, ELISE has no capability to understand the context in which keywords relate to each other. The conversational ELISE interface uses the same underlying search engine as the traditional interface.

---

used interchangeably with *virtual personality.*

III. METHOD

1. Overview of the Experiment and Hypotheses

This experiment compares two different types of interface models for a search engine: a traditional model and a conversational model (ELISE). The traditional interface resembles the interface of many Boolean search engines, while ELISE mimics a conversation between a virtual personality and a user in a chat window. Both interfaces share an underlying Boolean search engine. Users are asked to complete information-retrieval tasks using the two interfaces. I measure performance on several metrics and measure user satisfaction through surveys. Building on the conclusions of previous studies, the null hypotheses for the study are listed in Figure 5.

Figure 5 – Null Hypotheses for the Study

1. The conversational interface is just as fast or faster than a traditional interface.
2. The conversational interface is just as accurate or more accurate than the traditional interface.
3. People will prefer to use the conversational interface.

2. General Design of the Experiment

A. Experiment Summary

Users are asked to complete four information retrieval tasks (task sheet) while using one of the interfaces. They complete a similar task sheet when they use the other interface. By having each person use both interfaces, each person acts as their own control. I counterbalance the results by using two groups: one group with people using the traditional model first and the other using the NL model first. I measure the time to complete each task and the accuracy of the user's answers. To supplement the information of time efficiency, I also record the number of documents a user opens during each task and the number of searches submitted for each task. I also administer user satisfaction surveys asking the participants for their thoughts on: the accuracy of the interface's ability to identifying their keywords, the usability of the interface, the performance speed, and whether they liked the interface. The conversational interface evaluation also asks whether or not the virtual personality is believable. A survey asking the participants to compare the interfaces against each other is also administered. In addition to the surveys I record the user's typed input for both interfaces so that I have some context for what people did during their interactions. The data collected for performance and satisfaction determine the conclusions for this study.

B. Experimental Format for Participants

The experiment is divided up into 3 phases. Phase 1 consists of the participant filling out the pre-screening questionnaire. Phases 2 and 3 are when the participant attempts to complete

task sheets using one of the search engine interfaces. The experiment requires two groups of people. Group A uses the traditional interface first, and Group B uses the conversational first. Since both groups use both interfaces, each person acts as their own control. This study has 14 people split into two groups of 7. Both groups are similar to each other.

C. Experimental Format - Phase One

The participant fills out a prescreening questionnaire through a web browser in this phase. This questionnaire has questions about their general computer experience. The survey asks whether they have used Adobe Photoshop and whether they understand Boolean algebra. When the people finish the questionnaire and submit their responses, a program records their response and gives them a rating of beginner, average, or expert. This immediate scoring allows me maintain the experience similarity when choosing which group the person will belong to.

D. Experimental Format - Phases Two and Three

During these phases, the user attempts to complete the task sheet by using one of the search engine interfaces. If they are in group A the participant uses the user-driven interface first. If they are in group B they use the virtual personality. Regardless of the group, users have the same task sheet for this phase. Data is collected through several means: the task sheet, the videotape, the search queries, and the user-satisfaction surveys. Phase 3 is the same as Phase 2, except that the participant has a different task sheet and uses the other interface. Additionally the user receives an additional survey at the end of Phase 3, asking the person to compare the two interfaces.

E. Interface Designs in General

Shown below in are screenshots of the two interfaces: traditional (Figure 6) and conversational (Figure 7). (Larger screenshots for the two interfaces are provided and discussed in later sections.) The common elements between the two interfaces listed in Figure 8.

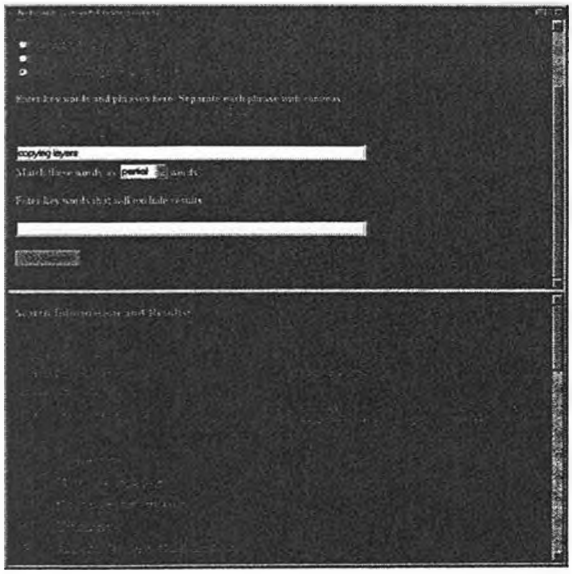


Figure 6 – A screenshot of the traditional interface used for this experiment

Figure 7 – A screenshot of the conversational interface used for this experiment

Figure 8 – Common Elements of Both Interfaces

- Both have user-driven models,[10] such that the user types in a search or sentence and the software responds to the user's input.
- Both have the same underlying Boolean search engine that displays the search results in a window section separated from the window used for input. The interfaces identify and send the search parameters to the search engine through the Common Gateway Interface (CGI[11]) protocol.
- Both interfaces make use of GUI components and layout through an Internet web browser. Users also open document windows by mouse-clicking a hyperlink.
- Both interfaces use the same background and text colors.

The underlying search engine has several input parameters for specifying a search. Although both interfaces have GUI components, the traditional interface makes wider use of GUI components to control these input parameters. The natural language interface does not use any GUI components to control these parameters. The NLI also has a virtual personality, and the recent dialog between the user and the software is displayed in a window. This recording and displaying of the ongoing dialog is similar to the interfaces for electronic chatrooms on the Internet.

F. Data Collection in General

I collect data through a task sheet, videotape, query recordings, and user satisfaction surveys. The task sheet includes the information retrieval tasks that users complete. The videotape records the computer monitor and the workspace area of the participant. Query

[10] If the virtual personality could initiate dialogue without any input from the user, then the interface would be a mixed-initiative interface.

recordings are input parameters for each search that is submitted to the underlying search engine, regardless of which interface they use. The user satisfaction surveys ask the user to evaluate their experience with the search engine.

G. Data collection - Task Sheets

The answer sheet is physically one piece of paper. (A blank answer sheet appears as APPENDIX A.) The paper lists question numbers and blanks where the participant is to fill out the answers of the task questions. The actual task questions are viewed through a web browser and linked together as web pages. The order of any series of task questions is as follows in Figure 9:

Figure 9 – General Format for a Task Set

Cover page: The user is told that there are 4 questions and to take their time completing the task sheet.

Question 1: A simple question asking the user to determine whether or not a particular document has a picture in it.

Question 2: A question that asks the user to find a document title that would explain how to do a particular task. The user is told that they do not have to read this document, and they can assume that the title explains the contents of the document.

Question 3: The question asks the participant to find the maximum number for some aspect of the Adobe Photoshop program. They are also asked to list the document title where they found this information.

Question 4: The user is asked to find a document that explains how to do a particular task. They are given a visual example of what they need to find. In order to answer this question, the user will need to read documents.

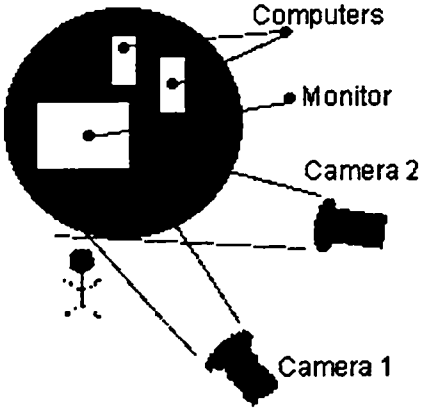
End page: The user is told they have finished the task sheet.

[11] Common Gateway Interface (CGI) is a very common method of having programs interact with each other over the

The purpose of requiring the user to write answers on the task sheet is to help evaluate the accuracy of a person's searches. Although a person may write the wrong answer on the task sheet, they may have opened the document that contains the answer. In such a case a person gets partial credit.

H. Data Collection - Videotaping

Two video cameras are in the room to record the workspace area and the computer screen. Both video cameras record to the same videotape. The camera aimed at the computer monitor is the primary picture in the recording. The workspace appears as a smaller picture area over the primary picture area. Figure 10 and Figure 11 show how the cameras map to the videotape image.



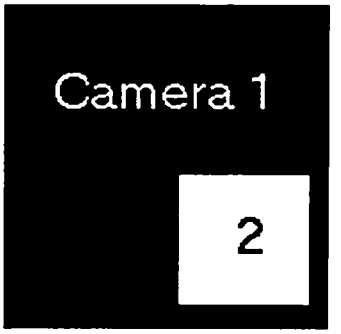


Figure 10 (left) – Layout of cameras and work area. Camera 1 focuses on the computer monitor while camera 2 focuses on the workspace area.

Figure 11 (right) shows that the images for Camera 1 appears as it would on a television screen (black area), while the images for Camera 2 appear in a smaller are (white box) over a portion of Camera 1.

Internet. Many of the search engines on the Internet use CGI for query submission.

The purpose of recording the workspace area is to give some context to what the person is doing. For example the person might not be typing or moving the computer mouse because he or she is writing the answer to the previous question. The purpose of videotaping the screen is to record the time it takes for each person to answer each question. Each time a user goes onto the next task sheet question signifies the person is starting a new question. In addition to informing me when the user starts a new question, the videotaping also informs me of the documents the user chooses to read. This is especially helpful in grading. If the user opens a document that would answer a task question but writes an incorrect answer, I award the user partial credit. The partial credit grading policy is important because some people may not necessarily be interested in doing the task sheet. Such people may not read documents as carefully as they would if the task was truly important to them. On the other hand, some people may think a particular document properly answers a question when the document does not provide the correct answer. The videotaped sessions keep a record of everything the user reads, enabling me to properly grade the task sheets.

I. Data Collection - Query Recordings

In addition to videotaping, the user's typing gets recorded. The search queries submitted through the traditional interface are recorded and stored in a file by the computer. The dialogue between the user and the virtual personality is also recorded. While these recordings do not necessarily provide data, they do provide insight into the kinds of things people did during their interaction.

J. Data Collection - User Satisfaction Surveys

After using a particular interface each participant fills out a four-question interface evaluation survey. The first three questions asked about accuracy/relevancy, speed performance, and usability of the interfaces. Each of these questions can be answered with a rating of –1, 0, or 1, where negative and positive numbers imply negative and positive responses respectively. The last question asked the users if they liked the interaction and their reasons for liking or disliking. At the end of phase 3, the user is also given a comparative-survey. The first three questions ask the user to compare the interfaces on accuracy/relevancy, speed, and usability. The last two questions ask which interface they would prefer in a workplace environment and which they would prefer in a non-workplace environment. In addition to these questions, the participants are also asked if they thought ELISE had a believable personality.

3. Participants, User Tasks, and Data

A. Participants

The general questions that were used for determining valid participants are as follows: *Are you comfortable using a computer while being videotaped? Do you know how to use a web browser? Have you used an Internet search engine before? Do you know about Adobe Photoshop?* These questions were asked informally before the person showed up to participate.

The most important requirement for a participant is that the person should not know anything about the Adobe Photoshop software. Hence the subject pool of participants is the domain of people who know how to use a web browser but know nothing about Adobe Photoshop. The choice to use people without such domain knowledge rests on the fact that those who do know about Photoshop may uniformly find their documents in such few search attempts that little is learned about the interaction between the user and the interface. People who do not know about Photoshop will more likely need to make more search queries, thus allowing for more insight into the interaction. The requirement for some web browser experience is so that the participants are generally comfortable working with hypertext environments. People who have used other computer graphics software are not excluded from participating in this experiment. Participants need not be regular users of web browsers and search engines. Variety in expertise should give some range for performance metrics. Being comfortable while being videotaped is important because two video cameras are in the room during the experiment. Although the cameras are not pointed at the person's face, a person's general uneasiness with cameras may cause worse performance.

The target audience for this experiment is people between the ages of 18 and 25, although people under the age of 18 and over the age of 25 can participate. The targeted age group has a lot to do with college campus location of the experiment. Gender is not an issue because people are not put into any kind of gender-based or gender-associated role. The only important aspect of gender is that both groups have the same (or similar) number of both genders.

The recruiting goals for this experiment are to have 14 people of mixed backgrounds that can be split into two groups of 7 while maintaining general similarity between groups. To help maintain similarity participants are asked to complete a screening questionnaire that asks them

about their computer experience and age. When they submit the form, the computer responds with a general rating indicating whether a person is a *infrequent*, *regular*, or *intensive* user. (The screening questionnaire appears in APPENDIX B.) Total scores for the questionnaire can range from 0 to 46. People scoring between 0 and 15 are infrequent users. People scoring from 15 through 35 are regular users. People scoring above 35 are intensive users. Generally speaking people who use a computer regularly will score higher, especially if they understand Boolean algebra. Since people can have different interpretations of how well they do Internet searches, expertise in doing search engines only carries moderate weight. Regular use of web browsers carries more weight than regular use of other software programs. Since Adobe Photoshop may have similarities to other graphics applications, the scoring is slightly higher than that of other Internet and office applications.

The user statistics and the breakdown of groups are shown in two tables. The first table (Figure 12) gives information on which group the person belonged to, their score on the screening survey, their experience rank (0 is infrequent, 1 is regular, and 2 is intensive), age, and gender. The second table (Figure 13) summarizes the composition of both groups.

USER ID	GROUP	SCORE	RANK	AGE	Gender
07011230H	A	8	0	1	F
07081000M	B	0	0	1	M
07151600T	B	42	2	2	M
07181900M	B	27	1	1	M
07181515V	A	44	2	2	M
07201430J	B	24	1	1	M
07211130M	A	33	2	1	M
07221130A	B	24	1	1	F
07231430J	A	27	1	1	M
07291400K	A	27	1	1	F
07291530O	A	17	1	1	M
07301300-	B	40	2	1	M
08011300K	B	26	1	1	F
08011700K	A	27	1	1	F

Figure 12 – Individual participants and their screening survey score, rank according to score, age, and gender.

	GRP A	GRP B	TOTAL
TOTAL IN GROUP	7	7	
AVERAGE SCORE	26.1429	26.1429	
INTENSIVE USERS	2	2	4
REGULAR USERS	4	4	8
INFREQUENT USERS	1	1	2
TOTAL UNDER AGE 18	0	0	0
TOTAL BETWEEN 18-25	6	6	12
TOTAL OVER AGE 25	1	1	2
FEMALES	3	2	
MALES	4	5	
TOTAL PARTICIPANTS	14		

Figure 13 – Summary composition of groups A and B according to the data in Figure 12.

Both groups appear equal in all ways with the exception that group A has one more female and one less male than group B. Both groups have 7 people with an average score of 26.1429 on the screening questions. Each group has 2 intensive users, 4 regular users, and 1 infrequent user. Nobody surveyed was under the age of 18, but both groups had 1 person over the age of 25. Everybody else was in the 18-25 year-old age category.

B. User Tasks in Depth

As mentioned previously, the format for the experiment is divided into 3 phases. Completing the previously described computer experience questionnaire completes the first phase. The second and third phases provide the bulk of the data gathered for this experiment. In later phases people use one of the interfaces to complete a task sheet. People in group A use the traditional interface first, and people in group B use the virtual personality interface. The tasks for each phase remain constant. Hence people will always start with Task Set 1 regardless of which interface they will use. They will do Task Set 2 in the third phase. By combining different task sets with different interfaces I will prevent the data for each interface from being biased. This is especially important because some people may complete one set of tasks faster than the other. If one task sets turns out to be easier to do than the other, the data collected for both interfaces reflect the easier nature of one of the task sets.

Participants are always also told several things when they start using an interface. This is to ensure that each person is aware of the basic aspects of the experiment. When participants begin phases 2 and 3, they are always told the following listed in Figure 14.

Figure 14 – Information Told to Each Person Before They Use an Interface

- The task sheet has 4 questions. It should take about 30 minutes to complete.
- They should take their time and make a sincere effort to complete it.
- They are not penalized in any way for wrong or incomplete answers.
- If they feel like they cannot answer the question, they can skip to the next question.
- They cannot go back to a previous question once they have left it.
- They should close windows after they've opened them because opening a currently open document from the search results window will not properly display the document.

Users who are using the traditional interface are told that they are using an interface similar to the search engines on the Internet. They are not told how to use the search engine. Users of the virtual personality are told the following things in Figure 15.

Figure 15 – Additional Information Told a Person Before They Use ELISE
- The interface is made to be able to converse and understand plain English.
- They can attempt to start a conversation with it if they wish.
- They are informed of the horizontal bar that signifies the end of ELISE's dialog.

After about 30 minutes or whenever the participant finishes the task sheet, the participant is given a user satisfaction survey. After completing the survey, the participant begins the second task set using the other interface. After 30 minutes or at time of completion they get another user satisfaction survey to evaluate the other interface. They also have an additional questionnaire asking them to compare the interfaces against each other. After this they are paid, and they leave.

C. The Background Scenario

The scenario that I have attempted to replicate for this experiment is *an information search and retrieval scenario on the Internet.* Domain specification further refined the ideal scenario to *an information retrieval exercise on a subset of the Internet*, where the subset is the documents dealing with Photoshop. I wanted the positive aspects of the Internet, without the problems of distracting advertisements, nonexistent hypertext links[12], and slow response time. However I chose to keep a particular inefficiency of many Internet search engines: limited accuracy of many searches. Many search engines on the Internet return document listings that are

[12] Some places on the Internet have hyperlinks to other places that no longer exist or maintained.

not necessarily relevant because those engines work with databases measured in gigabytes[13] of information and thousands and perhaps millions of documents. I wanted to observe how people might refine searches, so I chose replicate the limited accuracy of search engines. These decisions led to *an information retrieval scenario that used the hypertext documents of the Internet, the general designs of Internet search engines and Internet chatrooms, and the limited accuracy of the search engine.* The actual information retrieval tasks can be divided as follows: two questions asking the user to find a detail within a document and two questions asking the user to find information on how to complete a task.

[13] There are 1000 megabytes in a gigabyte. Many search engines work with hundreds of gigabytes. The data collection I worked with was less than 15 megabytes.

D. The Foreground Scenario: Task Questions and Performance Grading

Both task sheets mirror each other in how they are graded. Below is a summary explanation of how questions are scored in terms of completion time, number of documents opened, and number of searches submitted, and accuracy of recorded answers. (Actual task questions for *Task Sheet 1* and *Task Sheet 2* can be found in APPENDIX C and APPENDIX D, respectively.) Each question has its own accuracy table although all questions are graded on a range from 0 to 4 points, with four points being full credit. Since the goal of the experiment is to compare how well people find particular documents, time efficiency is weighted towards finding a document that answers the question. The starting time for a task is the time when the user begins a new question. The ending time is the time when the user locates a document with the correct answer. Accuracy grading gives a partial credit score of 3 points for finding the correct document but answering incorrectly.

E. Task Question One

This question simply tests whether a person can find a document if they are given the actual document title. The person is asked if a picture appears in the document. They can circle *yes* or no. The question is intended to be very easy. Accuracy scoring is summarized in Figure 16 below.

Figure 16 – Accuracy Scoring Table for Question 1 of Both Task Sets

4 points	User opens the specified document and answers correctly
3 points	User opens the right document but answers incorrectly
1 points	User opens an incorrect document that has one of the keywords in the title
0 points	User does not open document

F. Task Question Two

This question tests whether a person can find documents about a particular subject if they are given the subject words. Task set 1 has the subject of copying layers. Task set 2 has the subject of exporting an image. The user has been told that they can assume that document titles properly explain the content of the document. This aspect of the question reduces the problem of differing "reading times" for differing but acceptable documents. Some people may choose to open documents and read them to make sure they match the criteria. Other people may simply wish to list the documents without reading. By explicitly stating that people do not need to open any documents, people are more likely to not open documents. Accuracy scoring is as described in Figure 17.

Figure 17 – Accuracy Scoring Table for Question 2 of Both Task Sets

4 points	User lists a document title that contains both keywords.
2 points	User lists a document title that contains one of the keywords.
0 points	User lists a document title that contains none of the keywords.

G. Task Question Three

This question asks the user to find specific information about a general subject. No document title should reference the specific information, but the general subject matter should relate to the document title. Task set 1 has a general subject of *layers palette* and a specific subject of *maximum number of active layers*. Task set 2 has a general subject of *indexed-color palette* and a specific subject of *maximum number of colors*.

Post Data Analysis Note for Question Three

One problem about this question is that many people have difficulty finding the specific information due to the limited number of places where the information could be found and the fact that many people simply did not read documents carefully enough. The 3-point partial credit applies strongly to this question for both task sets. Since I need to

extract user error from software error, the time to complete this task is measured with the time when a user finds the document with the correct answer. Regardless of interface, most people found the proper documents, but they did not notice the document had the answer. It would be erroneous to mark time inefficiency against the interfaces because the interfaces have no control over how well people read documents. However the interfaces do have control over the documents listed in a search. Accuracy is graded on the following criteria shown in Figure 18.

Figure 18 – Accuracy Scoring Table for Question 3 of Both Task Sets

4 points	User opens a document that contains the complete answer after doing at least 1 search and answers correctly.
3 points	User opens a document that contains the complete answer after doing at least 1 search and answers correctly, but does not list the correct answer.
2 points	User opens a document that deals with both the general and specific topics but doesn't answer the question.
1 points	User opens the document with only the general or specific topic.
0 points	User does not open any documents that are about the related topic.

People will get better searches if they switch the domain to the helpfiles. People will also get better searches if they use the whole-word matching method while using the general subject keywords.

H. Task Question Four

This question is very similar to question 3, except it has a graphic that shows an example of a final product that can be achieved using Photoshop. The user is asked to find a document that details how to create the graphic. No document title should reference the specific information, but the general subject matter should relate to the document title. The general subject for Task Set 1 is *create text*, and the specific subject is *fire* or *fiery*. The general subject for Task Set 2 is *add shadow,* and the specific subject is *standing person*. Accuracy scoring is as described below in Figure 19.

Figure 19 – Accuracy Scoring Table for Question 4 of Both Task Sets

4 points	User opens a document that contains the complete answer after doing at least 1 search and answers correctly.
3 points	User opens a document that contains the complete answer after doing at least 1 search and answers correctly, but does not list the correct answer.
2 points	User opens a document that deals with both the general and specific topics but doesn't answer the question.
1 points	User opens the document with only the general or specific topic.
0 points	User does not open any documents that are about the related topic.

People will get better searches if they switch the domain to the tutorials. People will also get better searches if they use the partial-word matching method while using the general subject keywords.

Although the user has the ability to change different parameters of the search engine, the user is not required to change them. Both interfaces have the domain and word-matching method set to provide the widest range of results. Even though people will get better results by using different parameters, no suggestions are made within the question to do so. The traditional search engine does not make any suggestions for refining a search, but ELISE sometimes makes suggestions. However ELISE's suggestions are not derived from any knowledge of the task questions. Sometimes she will offer to modify a query and do another search. This does give the natural language interface a little more functionality than the traditional interface, but such an enhancement is appropriate for personality that wants to be helpful. In short words, ELISE will offer suggestions, but she does so without any assurance that those suggestions will actually help.

I. User Satisfaction Surveys

User satisfaction data is collected through user satisfaction surveys. The goal of these questions is to give some indication of user satisfaction in the areas of accuracy/relevancy, speed, usability, and enjoyment. The survey has one question for each of these performance areas. Those four questions can easily be scored using –1, 0, and 1, with positive numbers indicating positive feedback and negative values indicating negative feedback. The actual survey questions can be found in APPENDIX E and APPENDIX F. In addition to the four questions asked on both

interfaces, the survey for evaluating ELISE asks the participant for their thoughts on whether or not ELISE has a believable personality. This question is not a performance question, but I ask it so that I can properly assess whether I was successful at creating a virtual personality.

I also administer a comparative survey, asking the participants to compare ELISE against the traditional interface along the parameters of accuracy/relevancy, speed, and usability. I also ask for the person's preferences for an interface for both workplace and non-workplace scenarios. These five questions can be found in APPENDIX G. Responses to these five questions can be scored as –1 for a negative response, 0 for an undecided or neutral response, or 1 for a positive response. Although the process of having people fill an evaluation questionnaire for both interfaces and then asking them to compare the interfaces may seem redundant, asking the user to fill out the evaluations before the comparative survey gives context for why people may ultimately prefer one interface more than the other. The last two questions on the comparative survey ultimately determine whether or not people will prefer to use the conversational interface.

4. A Closer Look at Software Design

A. Layout of the User Interfaces

This experiment evaluates two user interfaces with one underlying search engine. The traditional interface derives its general design from various search engines on the Internet. Users have the following in the traditional user-driven interface. Figure 20 lists the layout characteristics of the traditional interface.

Figure 20 – List of Traditional Interface's Layout

- a textbox where they type their keyword phrase
- another textbox where they can type words used for exclusion
- radio-buttons[14] for selecting their domain
- a selection box[15] for selecting a word-matching method
- a button for submitting their search
- a dedicated section of the screen where the document results appear

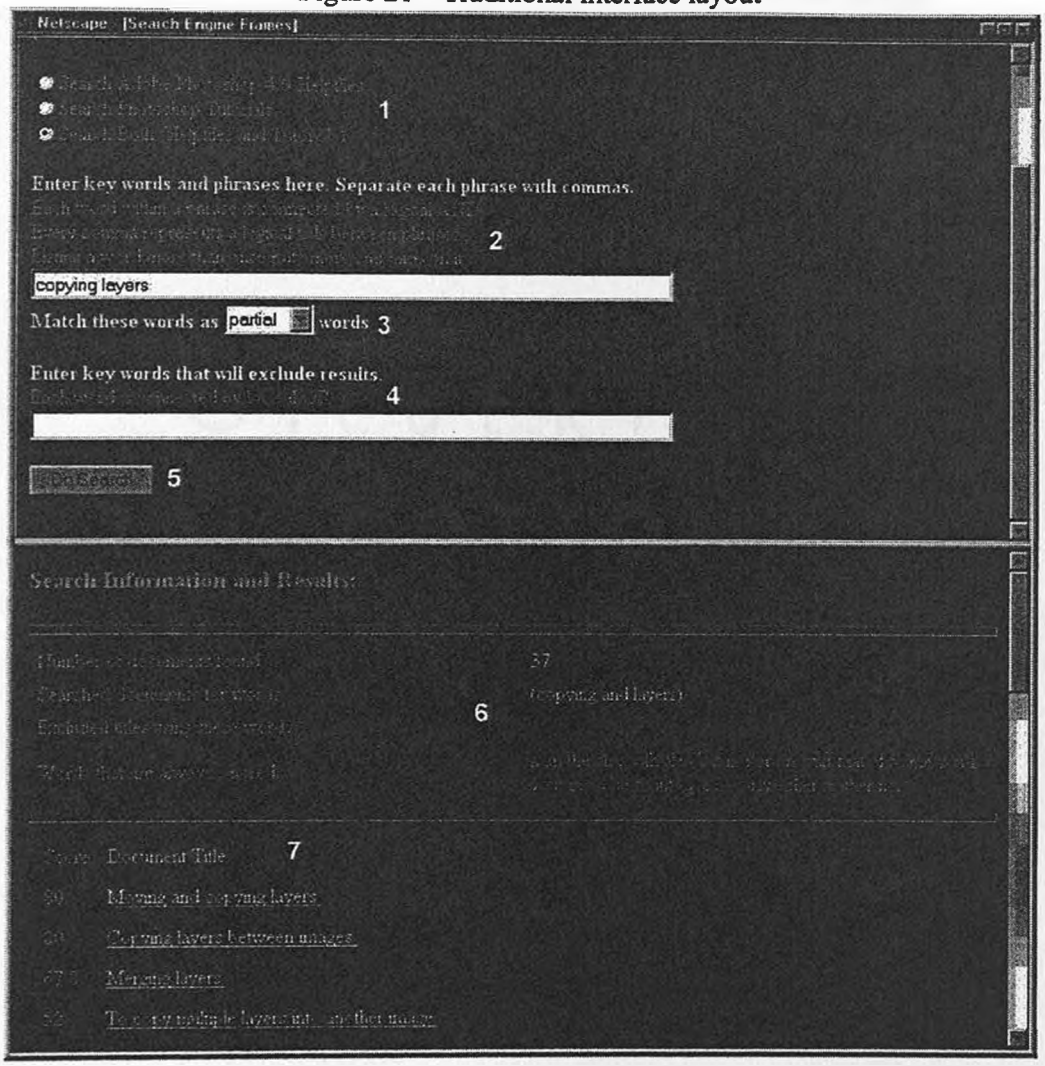
When designing the interface for the search engine with a virtual personality, I used the general design of Internet chatrooms[16]. A textbox is provided for a person to type in a sentence or phrase, and their sentence appears in a *chat-window* along with everybody else's dialogue. Since the user is only interacting with the software, the chat-window contains the most recently said things by both the user and the virtual personality. In the same manner as the traditional interface, the results of search queries appear in a separate section of the screen. Please see Figures 21 and 22 for a closer look at the layout of these interfaces.

[14] Radio buttons are a series of buttons that represent a series of options that a user may select. The user is only entitled to choose of one of those options, so each button *radios* the other button when it is selected. All other buttons acknowledge the selection, thus insuring that only one option is chosen at any particular moment.

[15] Similar in intent to a radio-button, a selection box simply allows selection of one of the items in a pull-down menu.

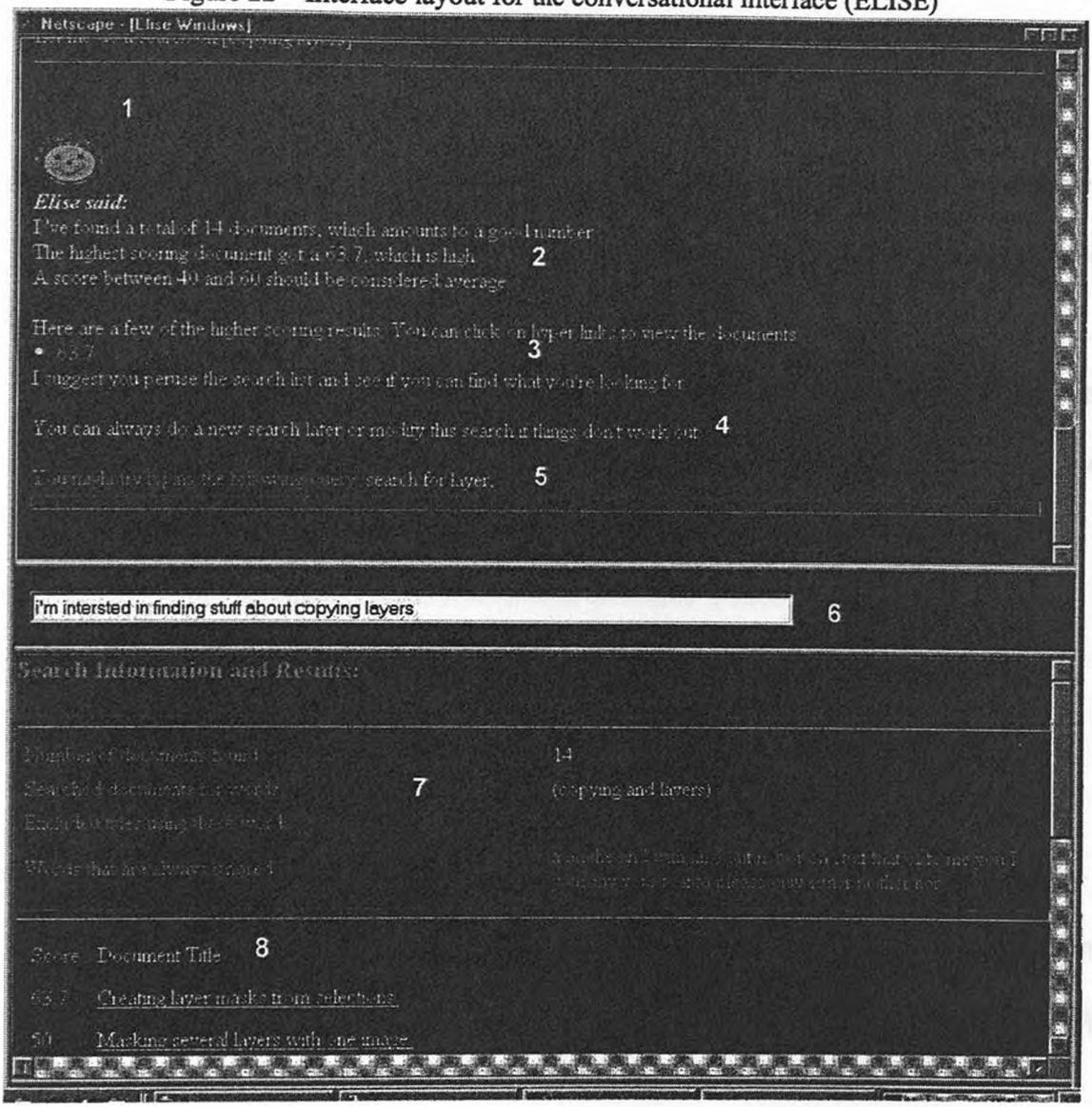
[16] Internet chatrooms are places where people can communicate by typing sentences to each other. A number of modem-based bulletin-board systems (BBSs) have chatrooms that are designed much like those of the Internet.

Figure 21 – Traditional interface layout



1. User selects the domain for the searches using the radio buttons
2. User types the desired keywords in the textbox. Information regarding the search engine's method of interpreting keywords is printed above the textbox.
3. User selects the word matching method with a selection box. The user can choose either to match the keywords as whole or partial words.
4. User types in words for exclusion. Words that appear in the document title that also appear in this textbox are excluded from the results.
5. User clicks this button to submit a search.
6. The search engine provides feedback on the search, including the number of documents found, the interpreted Boolean expression, words used for exclusion, and keywords that are ignored by the search engine.
7. The resulting documents found by the search engine appear in a list sorted by score. The user can scroll this portion of the window to see more results.

Figure 22 – Interface layout for the conversational interface (ELISE)



1. A graphic of ELISE appears here. She has 3 basic facial expressions, her standard smile, a smile with dimples, and a frown. The color scheme between the expression remains the same.
2. ELISE summarizes the findings of the search engine and comments on the reliability of the highest scoring document.
3. ELISE will have hyperlinks for the three highest scoring documents with scores over 60.
4. ELISE will make any comments she thinks are relevant to improve the search here.
5. ELISE sometimes suggests a search phrase that might yield better results.
6. The user types what they would like to say to ELISE in this textbox.
7. The search engine summarizes the results of the search in this area.
8. Documents found by the search engine appeared in a list sorted by score. The user can scroll this part of the window to reveal more documents in the list.

B. The Underlying Search Engine

The underlying search agent is written in the language Perl and has a Common Gateway Interface, much like any CGI script found on the Internet. The search engine has three distinct phases: evaluating input parameters and searching the database, and displaying the resulting list of documents.

When a user submits a search query, the interface chooses gathers the input parameters and sends them to the search engine through the CGI protocol. For the traditional interface the process of gathering these input parameters is quite straightforward as each GUI component controls one aspect of the input. The conversational interface is more complicated as the interface has to examine a plain English sentence and extract the proper keywords from the sentence or from the software's memory.

When the input is gathered and sent, the underlying search engine prepares all of the information needed to do the search. The search engine determines a Boolean expression, the possible variants of each keyword, and the files that are to be searched. The search engine then attempts to match the Boolean expression with the documents using a two-tiered search strategy.

There are two tiers of searching. The first tier of searching compares the Boolean expression against the document title. If the Boolean expression matches properly with a document title, it gives it a score. The search engine keeps a sorted list of the highest scoring documents in case of a second tier search. If seven or more documents are found in the first tier, the list is sorted by score, in descending order, and displayed to the user.

If fewer than seven documents are found, then a second tier search is started. During the second tier the search engine examines the actual text of the thirty highest scoring documents of the first-tier search, regardless of whether the document title properly matched the Boolean expression. The program examines the first 120 lines in the document, giving a score to each line in the same manner that it did for the document title.

C. A Quick Note on Searching and Boolean Expressions

The following is a simple example of how documents might be found by the search engine. Suppose the search engine extracted the following Boolean expression from some input: (cats AND dogs) OR (cats AND mice). The following titles will match.

Cats and dogs together again,
Cats like mice,
Mice fear cats and cats fear dogs

The titles match because the Boolean expression is attempting to find documents that deal with *cats* and *dogs* or documents about *cats* and *mice.* The first document is cats and dogs. The second is about cats and mice. The third is about cats, dogs, and mice. All three document titles fit the criteria of the Boolean expression.

However there are instances when a Boolean expression is too strict, making the document list very short. At the same time, the document may contain the needed information. For example the Boolean expression (cats AND fish AND dogs AND birds) requires that all of those words be present in the document title during a tier one search. Hence, a document like *Cats like fish and dogs* would not match the Boolean expression, but it might score high because three of the four words are present in the title. During a tier-two search situation, such higher

scoring documents that did not meet all the requirements are examined more in-depth. The search engine actually reads some of the document's text in order to determine whether or not the document might fit the criteria.

5. Interfaces in Depth

A. Traditional Interface in Depth

Although the underlying search engine does all of the searching, the input sent to the search engine rests largely upon the user interfaces. As mentioned earlier the traditional interface uses graphical user interface components. Each of the GUI components directly controls a particular aspect of the search query, which makes the traditional interface quite straightforward and easy to use. Furthermore the traditional interface also has basic information on the input parameters. By default this interface is set to search both the helpfiles and the tutorials while using a partial-word matching method. The search engine's basic execution cycle when using the traditional interface is as explained in Figure 23.

Figure 23 – Execution Pattern for the Traditional Interface

1. User types in the keyword phrase and any words for exclusion in their respective textboxes. The user also chooses/changes the word-matching method or the domain for the search.
2. User submits the search query. The information recorded by the user-driven interface is sent to the search engine through via CGI.
3. The search engine performs a search using the information in the search query.
4. The search engine displays a list of scored document titles in a different window[17].
5. The pattern repeats at step 1.

B. Conversational Interface in Depth

The virtual personality used for this experiment has been titled ELISE, which stands for Extremely Limited Intelligence Search Entity. (Although the name may seem self-degrading, the humility in her name fits her personality.) ELISE can actually explain the acronym for her name, if anybody asks. People can also ask her about her preferences: *What do you like? Do you like oranges? Do you hate anything? What music do you listen to?* Of course her main purpose, as she would say, is to "help you find documents about Photoshop." One could tell her a variety of things. Some examples of dialog are listed in Figure 24. Each bullet begins a new series of statements that can be said sequentially. Italicized words, phrases, and letters indicate that keywords can be substituted for the listed words.

[17] The user driven interface has its own dedicated section for displaying the results from the search engine. In more technical terms, the search query submission and the search query results have their own web browser frames.

Figure 24 – Examples of Statements Understandable by ELISE

- Find stuff about *cats, dogs, and mice*
- I'm interested in finding out about *chickens*
- What can you tell me about *holiday travels in Jamaica*
- Search the helpfiles for *pictures, pens, and ink* but ignore stuff on *pencils*
- Switch to whole word matching method and find stuff about *cascaded images*
- Search for [*printer resolutions*] but not {*monitors*}[18]
- I need help with *drawing a pig*
- I want to know about *dogs and cats*
  And *mice* too
  but not *rats or pidgeons*
  or *flies*

In short ELISE supports the full expressive power of the traditional interface and provides a little more. Her help dialogue is much more extensive. She can provide more examples about different kinds of searches, with more in-depth explanations about some of the options of the search engine. She tries to offer tips on how to modify searches after examining the most recent search. She sometimes offers to submit a modified version of the last search, in case the user does not find anything. Other time she suggests a query that she believes might be more efficient[19].

The interface with the virtual personality is much less straightforward in technicality. This natural language interface is created in such a way so that the virtual personality can perform and refine searches through a conversation. Please see Figure 25 for an example of a short interaction with ELISE. The formatting of the dialog has been slightly modified for paper.

[18] If for some reason ELISE has trouble figuring out keywords, the user can surround the intended keywords with [brackets] and surround the intended exclusion words with {braces}.
[19] If a user submits a query with more than three keywords, ELISE will attempt suggest a new query that contains certain keywords that are most relevant to Photoshop.

The natural language interface is implemented through the commercially available software, Neurostudio[20], and is available from Neuromedia, Inc. The Neurostudio program comes with its own programming language, GeRBiL[21] that can do complex forms of pattern matching[22]. Neurostudio can also remember basic things that have been said or done in previous portions of its interaction with a user, which allows for differing levels of anaphoric and elliptic capabilities. Neurostudio also has CGI capabilities that allow it to interface with the search engine in the exact manner the traditional interface did.

[20] Neurostudio was created with the expressed purpose of providing virtual personalities for use on the Internet. Neurostudio can interface well with other Internet components, such as search engines, which contributed greatly to my decision to use it.

[21] GeRBiL stands for General Robot Building Language

[22] Pattern-matching is one strategy for doing natural language interfaces. It works on the basis of examining a sentence for certain word patterns. When it finds a particular word pattern, the program responds to the pattern according to how it was programmed to respond. For example the word *find* might set the NLI into a mode for finding things.

Figure 25 – Example of dialog with Elise

Elise:

Hello. I'm Elise. My job is to help you find information on using Adobe Photoshop.
You can ask me things in plain English, and I'll do my best to understand you.
If you need any help, just type "help".
The examples I can show you are somewhat formalized, but I encourage you to try normal language with me.
I'm hoping to learn more about how people speak.
Are you interested in seeing some examples on doing searches?

User: hi

Elise:

Hi there.
Now that we've said hi, why don't you ask me to find something about Adobe Photoshop.

User: what's your name?

Elise:

My name is Elise. It's short for Extremely Limited Intelligent Searching Entity.

Although this chit chat isn't doing any real harm, I can't say that I'm doing anything useful. Why don't you have me find a document about using Photoshop.

User: would you search for things on copying layers, please?

Elise:

You used several words separated by spaces. This may not give you enough documents. If this happens you may wish to use more commas.
You will find the details to my search on [things copying layers] listed below.

Elise:

I 've found a total of 18 documents, which amounts to a good number.
The highest scoring document got a 111, which is very high.
A score between 40 and 60 should be considered average.

Here are a few of the higher scoring results. You can click on hyper links to view the documents.

- 111 Blending_Layers_with_the_Options_Palette.html
- 107.5 Drop_Shadow_with_Layers_Tutorial.html
- 106.3 Masking_Several_Layers_with_One_Image.html

I suggest you peruse the search list and see if you can find what you're looking for.

You can always do a new search later or modify this search if things don't work out.
You might try typing the following query: search for layer,

User: redo the search using tutorials

Elise:

I'll be searching the tutorials now.
Ok. I'll search using the current parameters.

Elise:

I 've found a total of 16 documents, which amounts to a good number.
The highest scoring document got a 63.7, which is high.
A score between 40 and 60 should be considered average.

Here are a few of the higher scoring results. You can click on hyper links to view the documents.

1. 63.7 Creating_Layer_Masks_from_Selections.html
2. 51.2 Drop_Shadow_with_Layers_Tutorial.html

You can look through the results. Maybe you can find what you're looking for.

You can always do a new search later or modify this search if things don't work out.
You might try typing the following query: search for layer,

C. How is ELISE Set Up?

ELISE determines a proper response to a user's statement by examining a long list of possible sentence patterns. She matches the statement to a pattern and attempts to figure out the details of that sentence. Sometimes the sentence may have a request to do multiple tasks, such as *switch to partial word method and search the tutorials for x, y, z too*. A key aspect to partner matching is the presence of certain words in a statement will direct the course of her response.

D. A Searching Example for ELISE

In the previous statement *switch* and *partial-word* appeared sequentially, telling ELISE that she should switch to the partial word matching method. The word *search* was in the statement, telling ELISE that there may be an upcoming keyphrase. She also sees *tutorials* shortly following search, telling her to use the tutorial domain. She comes across the preposition *for*, sees that this preposition shortly follows *search*, indicating that she can expect keywords to appear after the word *for*. She also notices that the word *too* appears sometime after *for*, so she assumes that everything between *for* and *too* is a keyphrase. So, she switches to partial word matching method and submits a search for *x, y, z* in the tutorial domain. She waits for the search to complete, and then examines the search query and the results of the search engine. Using this data, she attempts to make some suggestions for other types of searches. She might recommend the user to read the three highest scoring documents, or she might offer to submit a modified version of the search. She then awaits the user to input a statement. When she gets the statement, she repeats the process over again. The user may have asked for help on doing searches, so she

may end up giving feedback on doing efficient searches. How ELISE responds is very dependent on what the user types and what immediately preceded the user's statement: *Did the user do a search? Did ELISE already make a particular suggestion for this search? Did ELISE ask a question? Is the user responding to that question, or has the user switched to another topic? Is the user asking to find things that don't deal with Photoshop?*

E. ELISE Confronting Various Scenarios

In addition to doing searches and providing feedback on how to do better searches, ELISE can also respond to other statements that do not deal with Adobe Photoshop. Although she never volunteers such information on her own, she can respond to statements that may seem off-topic to her. For instance, if somebody asked her about her name, ELISE could choose to explain that her name means Extremely Limited Intelligence Search Entity. However she would immediately follow up with a statement encouraging the user to finish the task sheet, saying something like: *That's enough about me. Now let's go do that task sheet.* If somebody asked her about her preferences, she would say that she really didn't have any. Similarly if somebody asked her if she liked something in particular, she would say she did not know enough about the subject to really know. Again, she would follow up with a statement reminding the user to finish the task sheet. She even responds to negative feedback from the user, such as, *You're stupid*, and follows up with apologies, citing her lack of intelligence: *Sorry for any inconvenience. I'm not as smart as you.* ELISE can take as much verbal abuse as a person wants to type, and she will ever remain polite and apologetic.

F. ELISE's Personality and Anthropomorphism (Her Ability to Seem Human)

One of the most difficult things about creating a personality for ELISE is that she lacks many of the human qualities because ELISE is a software program. What was needed was a defined agent metaphor[23] suited to her job concept. Since her job was to be functional, I focused her personality towards being a functional but polite personality. She conveys very little emotion, other than polite statements such as: *you're welcome, no problem, all right,* and *thank you.* She is very apologetic, being fully aware of her limited intelligence, but she never resorts to angry or resentful tones. (This apologetic personality may be friendly and courteous to some but may be spineless and weak towards those who prefer more assertive personalities.) She *smiles*[24] when she does a search that returns a good number of documents and frowns when she does not do a good search. To develop her *functional search agent* metaphor, I bestowed her with two goals. She explicitly says her primary goal when she meets a new user, but she also has the goal of performing better than the traditional interface. Hence she attempts to be more helpful by providing more information to the user. Still fitting in with her metaphor she tries to make the information digestible by presenting it in smaller amounts. This prevents people from being overwhelmed with pages and page of unending help that she has. She also offers to submit a variation of a search for the user, so that the user does not have to ask for it. She also attempts to encourage a user to finish the task sheet whenever they stray off-topic, but she also lets users know that she talks about how she feels about Photoshop or the task sheet in an attempt to seem unbiased about her situation. Despite her unspoken goal, she never says anything that would degrade the traditional interface or make herself seem better because software that praises itself

[23] An agent metaphor is description of what the software agent is trying to be.

and criticizes other things generated negative feelings on the part of the user. (Nass, Steuer, Tauber, and Reeder, 1993)

6. Underlying Search Engine in Depth

A. Input Parameters in General

The underlying search engine has several input parameters that determine what documents it will attempt to find and exclude and how it finds those documents. Collectively the input for the search engine is also known as a *search query*. Input includes the keyword phrase, exclusion words list, word-matching method, and domain. Since both interfaces have the same underlying Boolean search engine, both interfaces have means to control the search queries. In the traditional interface the use has more control over this input because each GUI component is tied to an input parameter. The natural language interface has to formulate a search query based on its interpretation of what the user has said. Regardless of the interface, the search engine uses the input parameters sent from the interface to perform a search. The following is an explanation of the various input parameters.

B. Keywords, Grammar, and Interpretation

The *keyword phrase* is a phrase that translates into a Boolean expression of topics. The search engine uses the Boolean expression to find documents that deal with the topics in the

[24] The interface displays one of three pictures, as an indication of ELISE's mood.

keyword phrase. The keyword phrase has the following grammar and Boolean interpretations listed in Figure 26.

Figure 26 – Lexical Grammar of the Underlying Search Engine

- A *space* and a *comma* are as they are in the English grammar.
- A *series* is considered a set with one or more items in it.
- A *word* (or *keyword)* is a string of characters terminated by either a *space* or a *comma.*
- A *series* of *words* separated only by *spaces*, yields a *keyphrase.*
- A *series* of *keyphrases* separated only by commas, yields a *keyword phrase.*
- A *keyphrase* is always terminated either by a space or the NULL[25] character that terminates the entire keyword phrase that was used as input.
- Every *comma* represents a Boolean logical OR between *keyphrases.*
- Every *space* represents a Boolean logical AND between *words.*
- Every case-insensitive[26] *word* in the list of *ignore words* is removed from the keywords phrase. The *ignore words* list is not an input parameter as it is defined within the search engine itself. Words in the *ignore-words* list are very common so they are ignored whenever creating the Boolean expression. A complete listing of the *ignore words* is as follows: a, an, the, and, with, also, but, not, or, on, stuff, that, of, to, me, you, I, your, my, your, to, into, please, only, either, neither, nor.

[25] Effectively, the NULL character is marked by the unseen carriage-return that the user types.
[26] Case-insensitive refers to undistinguished treatment of capital or lower-case letters. The search engine does not distinguish between capitalized and non-capitalized letters.

As an example, the above grammar would evaluate a keyword phrase in the manner shown in Figure 27.

Figure 27 – Example Translation of Keywords Phrase into a Boolean Expression

- Input keywords phrase: cats and dogs or fish, but not bulls cows, the chickens, OR hens
- Translated into keyphrases: cats dogs fish, bulls cows, chickens, hens
- Translated into Boolean expression: (cats AND dogs AND fish) OR (bulls AND cows) OR (chickens) OR (hens)

C. Exclusion Words

The *exclusion words list* is a list of words that are used to exclude documents, whose titles have words that match with any of the words in the *exclusion words list.* For example, if a document's title was "Fred likes fish', and *fish* was one of the words in the exclusion words list, then the document would be excluded as a possible result for the search. If a word in the exclusion words list also appears in the keyword phrase, then the exclusion words list takes precedence. This means that exclusion words will always exclude documents that match with a word in their list regardless of whether or not the word appears in the keyword phrase.

D. Word Matching and Domain

After a Boolean expression is determined from the input, the search engine attempts to search its database for documents that match the Boolean expression. While the *domain*

determines which documents get searched, the *word-matching* method affects the strategy for how documents get matched. The *word-matching* method has two settings: *partial* or *whole.*

Partial word matching will tell the search engine match variants of a word, while *whole* word matching looks for the exact word. Figure 28 is a listing of how different kinds of variants may be generated for a keyword.

Figure 28 – Generating Variants of a Keyword

- Hyphenated words such as *information-retrieval* are split into their separate words. Each word is added to a list of variants for the keyword. Both the hyphenated keywords and the hyphenated words in a document are split for variants. Thus, words like *information* and *information retrieval* would match each other.
- Words that are a substring[27] of a *keyword* are considered variants as well, although these variants are not created and placed in the variant list. The *Perl* language has functions for searching substrings so there is no need to create additional words for the variant list.
- Other variants of the keywords include any word that can be created by dropping any of the following suffixes (-ives, -ions, -ings, -ive, -ion, -ing, -es, -ed, -s, -e) to the keyword. Thus if we take the keywords *formations,* then the variant would be *format* and *formation.* When attempting to match a word in with one of these variants, the word's substrings can also match. Therefore the word *formation* could match with the variant word *format* derived from the keyword *formative.*

If the search engine is using partial-word matching, and if it fails to match a document's word against a keyword, the program tries to match the keyword's variants against the document's

word[28]. Matching is case-insensitive so words like *format*, *FORMAT,* and *FoRmAt* are all treated the same.

If the method is set to whole word matching, then only hyphenated words are included as acceptable variants. Substring variants and other variants are not matched. Hence, *format* would only match with the word *format*, but *information retrieval* would still match with *information.* Word matching is also case-insensitive for whole-word matching.

The search engine works with a collection of HTML files about the Adobe Photoshop software. There are a total of about 10 megabytes of files taken from the Adobe Photoshop helpfile and various websites. Thus, the possible search domains are:

- Searching the Adobe Photoshop Helpfiles
- Searching the tutorials downloaded from websites
- Searching both the Helpfile and the tutorials

When a search query is submitted the search engine chooses a list of files that belong to the domains. The search engine then performs the Boolean search only on the documents in requested domain.

E. Searching and Scoring Documents

When the search engine receives the input, determines a Boolean expression, the possible variants of each keyword, and the files that are to be searched, the search engine then attempts to find documents. The algorithm for finding documents has two tiers. The first tier search attempts

---

[27] *Substring variants* are words that appear exactly in another word. For example, *reference* appears in the word *preference* and the word *aliasing* appears in the word *anti-aliasing.*

to find documents by examining the words in the documents' title. If a first tier search does not return at least 7 documents, the search engine will attempt a second tier search, which lets the search engine search the first 120 lines of actual text within a document. A second tier search will examine the 30 highest scoring documents of the first tier search. In the end the search engine sorts the list of documents according to a relevancy score, and the list is sorted in descending order and displayed to the user.

The algorithm for the search engine is shown in Figure 29. Recall that a Boolean expression takes the form of a series of keyphrases connected by a logical OR. Each word within a keyphrase is connected by a logical AND. *Substrings* are the words within words. For example the word *kitty* is a substring of the word *kittyhawk.* A *variant* of a word will remove suffixes like *-ion, -ing, -ed, -es,* and *–s* from words. Recall that substrings and variants are ignored if the user is using a whole-word matching.

[28] The document's word that the search engine is examining might be from the document's title or the document's article.

Figure 29 – Algorithm for Finding Documents

1) Take the keyphrases of the Boolean expression and put them in list: (@keyphrases)

2) Take the words of the document's title and put them in a list (@titlewords)

3) Examine a keyphrase by doing the following

   a. If a word ($tword) in @titlewords matches exactly with a word in the keyphrase, then add 2 to the score.

   b. Otherwise if the $tword matches a substring of a word in the keyphrase matches with the keyword, then add 1 point to the score.

   c. Otherwise if the word matches with a variant of a keyword, then add some fractional points. If there are x variants for a particular keywords, then the fractional points are worth $1/x$.

   d. If all the words in the keyphrase matched with words in the document's title, multiply the score by 20 and mark this document as a valid result.

   e. Otherwise if all of the words in the keyphrase did not match, go back to step 3 and analyze the next keyphrase. If no keyphrases matched then the document is excluded as a possible result and go to step 9.

4. Sort the results.

5. If there are more than 7 documents go to step 14.

6. If there are fewer than 7 documents, then start a second-tier search. Examine the document by comparing the first 120 lines of the 30 highest scoring documents against the each keyphrase. Score each line in the same manner detailed in steps 4 and 5. Every time a line matches the Boolean expression adds the score from the line to the score of the document.

   a. If the 30 documents have been searched then go to the next step. Otherwise compare the next document against the keyphrases.

7. Sort the results and take the 15 highest scoring documents.

8. Display the results of the search.

7. Data Analysis

The independent variable for this experiment is the condition. In this case the condition is the type of user interface being evaluated: the virtual personality interface capable of natural language or the traditional interface. The dependent variables for evaluation will be efficiency metrics and user satisfaction. Efficiency is measured through several criteria along each task question: time, accuracy, number of unique documents viewed, and number of searches submitted. Recall that each of these criteria have a grading scale, with a particular bias towards actually finding the document. User satisfaction is evaluated through the questions on the evaluation and comparative surveys. User satisfaction is the sum of the points associated with each response to the questions on those surveys.

The data analysis for this experiment uses a standard Analysis of Variance (ANOVA) over repeated measures. I will use the ANOVA method to analyze performance metrics: time, number of searches submitted, number of documents opened, and accuracy between both groups. ANOVA will determine the likelihood that the averages for each metric are different by comparing variation amongst groups with the variation within groups. ANOVA also gives a value for the probability (P-value) of making a wrong statistical conclusion. The ANOVA tests essentially ask the following question: *Is there a difference between two interfaces?* The null hypothesis for each of the performance is this: *There is no statistically noticeable difference between the averages for the performance metrics: time, number of searches submitted, number of documents opened, and accuracy when comparing the interfaces.* Please note that ANOVA is applied to each of the performance metrics; the tests do not compare these metrics together. If a P-value for a test is less than 0.05 (5%), then null hypothesis is less than 1 in 20 chance of being

wrong. Such a result indicates that the findings are considered statistically significantly and different. Conversely, a very large P-value implies a possibility of equivalence.

The ANOVA test is a standard parametric test used to test for a difference between two treatments. ANOVA over repeated measures assumes that every participant's results are independent of each other. In other words the performance of one person does not affect the performance of another person. Since both interfaces do not change in this experiment and since participants do not interact or discuss anything with each other, the experiment meets this criterion. ANOVA tests also assume that the data of each performance metric is distributed normally around its mean, with the standard deviations being within a factor of 2 between each other. Essentially this means that both interfaces have similar standard deviations for their averages. Since this experiment involves each person doing task sheets twice, I need to use the ANOVA over *repeated measures* test. In this case the repeated factor is the act of doing the task sheet twice. The ANOVA test also assumes that both experimental groups are randomized and similar in such a way that neither group is biased towards producing a particular result. Since group selection is determined by a computer generated score, this criterion is effectively met.

The equivalence test used for this experiment will be the parametric test as explained in the article, *Using Significance Tests to Evaluate Equivalence Between Two Experimental Groups*. (Rogers, Howard, Vessey, 1993) Instead of testing for a difference in the performance, it tests for an equivalency in performance. In this experiment this test is applied only when an ANOVA test gives a very high P-value (90% or higher) with the means and standard deviations appearing very similar between both groups.

The data analysis also uses the Wilcoxon Signed-Rank tests to analyze the user satisfaction data. This test is used because the data is ordinal. People who answered the user satisfaction questionnaire could choose from 3 possible choices. These choices map to individual values, with no possibility of a user choosing a value between the three designated values, thus indicating that the data is ordinal, as opposed to a direct measurement. The Wilcoxon Signed-Rank test is designed to test a hypothesis about the average person of a population distribution. In this case I am testing whether or not people thought the interfaces: provided accurate and relevant searches, were fast in doing searches, and were easy to understand and use.

The Binomial Single Proportion Test is also used for determining the interface that most people prefer to use. This test is used to test the last two questions on the *Comparative Survey* asking the user to choose an interface they would prefer to use in a workplace and a non-workplace situation. The Binomial Single Proportion test assumes two things: *For those participants, who preferred an interface, there is a fixed probability P that they will choose ELISE. All participants' judgements are independent of each other.* In plain terms this means that a probability of person choosing ELISE does exist and that the person's judgements are not affected by other people's judgements of ELISE. The null hypothesis for this test is as follows: *The probability of the average person choosing ELISE is 0.5 or 50%.* If the test indicates a probability higher than 0.5, then it means that people have a preference for ELISE. To calculate a confidence level percentage[29] we simply calculate 1-2*(1-P). The confidence level is a percentage indicating the reliability for the P-value in this test.

[29] Please see APPENDIX H for an explanation of why this formula works.

8. Issues and Justifications

A. Why does the traditional interface have a separate textbox for exclusion words?

While most of these traditional interface components are common in search engines[30], I admit that the additional textbox is not a regular part of many search engines found on the Internet. The decision to use a separate textbox for exclusion words rises from the fact that many Internet search engines have their own query grammar and syntax. Since many grammars exist and people may be influenced by the search grammars they use on the Internet, I avoided the *NOT* Boolean operator believing that it could contribute to usability problems. For example, the phrase below could be interpreted in two ways:

Phrase: cats and dogs and not fish and mice

Interpretation 1: look for cats and dogs but exclude anything about fish or mice

Interpretation 2: look for cats, dogs, and mice but exclude anything about fish

To help prevent ambiguity when using the *NOT* Boolean operator, I chose to give a separate textbox for the exclusion-words that would be used for excluding documents from results.

B. Why do the search engine results always appear in a separate portion of the window?

While most Internet search engines do not have a dedicated section of the screen for results, I choose to separate the search results from the area where the user types their input. The decision is motivated by the desire to make the division between the underlying search engine

and the interfaces more apparent. It is my intention to make the search results the only visible component to the underlying search engine. In addition to this division, splitting the screen also allows me to create two the interfaces so that their general layout and color remain the same between phases 2 and 3. (During one of the experiment runs, one participant commented that having the results in a separate window is convenient.)

C. What was the motivation for choosing Adobe Photoshop as the topic for the database? What is the size and complexity of the database?

Most of the difficulties in obtaining a large database are due to the fact that a number of people who had large collections of documents refused my requests to obtain a copy of their database. During my search for a database, I obtained permission to decompile the Photoshop user documentation from the Adobe Software Company. I also received permission from various Internet webmasters who gave me permission to use their Photoshop tutorials for this study. There are 890 documents in the database, with about 10 megabytes from the Photoshop helpfile and another 2 megabytes from the tutorials. Although the tutorials do not have hypertext[31] links between them, a number of the helpfile documents are linked through hypertext. The choice to use Adobe Photoshop is also motivated knowledge of Photoshop tends to be specialized and finding people who know nothing about the software is not too difficult.

---

[30] Textboxes are often used for keyword phrases, while radio-buttons and selection boxes are often used to modify the search engines behavior in finding documents.

[31] The *links* on webpages that lead to other webpages are hypertext links.

D. Why did you write your own search engine instead of modifying an existing One?

There are a number of efficient Boolean search engines that might seem like ideal candidates for the underlying search engine. However the biggest problem with most of them is that they are too efficient for a database collection of less than 15 megabytes. After trying one such commercial search engine I discovered that it would almost always narrow down the proper result as one of the top three. I also found that such a search engine did poorly when it received queries from the natural language interface. I decided that such an effective Boolean search engine could unjustly bias the results to the traditional user-driven interface. I chose to write my own search engine that could more easily adapt to ELISE's search queries while replicating the limited accuracy found on the Internet.

E. Why does ELISE not have more enhanced anthropomorphic abilities, such as speech and microphone input?

One could argue that technology has increased to such a level that a visually animated, speech-capable virtual personality can interact through a microphone for the dialogue. While this is true in many cases, I discovered some problems when considering such an advanced anthropomorphic design. I would have liked to a visually animated image to represent ELISE, but I did not have the technical means or the resources to acquire one that would work well with the underlying search engine. Speech capability definitely adds more of an anthropomorphic quality to ELISE, but preliminary testing showed that the ELISE talked much slower than people could read. I also discovered that a lot of software that converted speech to text required time for the software to adapt to the user, and the results prior to adaptation were not very good. Since the

time to complete a question is a performance metric, I chose to exclude the speech capabilities that biased time and efficiency against natural language.

F. Most of the effective natural language interfaces can understand the context of various words with an advanced lexical analyzer. Why does ELISE use pattern-matching instead of a more advanced process? Does ELISE qualify as a natural language interface? What kinds of anaphoric or elliptic capabilities does she have?

Many of the successful commercial natural language search systems can identify the context of words within a sentence and then search a presorted database that works in conjunction with the context-capable search engine. The problem is that such a scenario unfairly biases the natural language search engine towards better performance because a simple Boolean search engine has no way to make sure of context-capable databases. The decision to use a pattern-matching scheme rests largely upon what was available to me. The Neurostudio software, with its capabilities to interface with other components through CGI and generally simple programming language, made it an ideal platform for creating ELISE. She can carry a very limited conversation, and ELISE does attempt to keep the dialog within the bounds of completing the task sheet. People can ask her about other subjects, and she attempts to respond like a person without any real bias or preference. Using the evidence that anaphoric or elliptic capabilites can be replaced by recall-and-edit (Burton, Steward, 1993), I minimize the effort placed in ellipsis and anaphora, using them only to modify search results with phrases like: *add x,y,z to the keywords, but not a,b,c, now exclude x too,* etc. One limitation of ELISE is that she can only make one search request at a time. The recall-and-edit functionality is limited in that the user's last statement is automatically recalled only once. (The user-driven interface also has this

limitation. All of the query parameters automatically recall the most previous input.) The issue of ELISE's qualifications as a natural language interface is best determined by the findings of this experiment. If people complete their tasks with ELISE and like using ELISE, then such evidence is an indication of her natural language robustness.

G. How is ELISE a believable virtual personality?

While the true measure of ELISE's *believable* personality rests on the perception of the user, the theory used for building ELISE comes from methodology that recommends the use of personality quirks, competence, emotions, relationships, attitudes, norms, roles, goals, and robustness (Reilly, 1997). Part of the challenge in creating a believable personality was choosing the aspects that would not negatively affect the overall performance of the search queries. For instance, a personality that became displeased with a user and refused to submit searches or submitted inefficient searches in retaliation would be very inefficient. ELISE's basic personality stems from her goals: *find documents about Photoshop* and *perform better than her user-driven rival.* What immediately follows from this is that she has a relationship to the user. The user has become her client, and she is to help the user as best as she can. Thus her behavior is constrained to being courteous and respectful towards the user, understanding that she's not really in control of things. She has a habit of saying, "you're the boss" or "whatever you say" when she makes a suggestion and finds that the user did not take her advice. Although ELISE is competent enough to be part of this experiment, she is quite aware that she is not perfect in her understanding of the English language. Hence, she is cautious about how she approaches the issue of natural language queries, informing the user that she can, in fact, understand them. However, whenever she gives

help to the user, she assumes that the user wants the most functional information, and so she tries to give simple examples and works her way up to more advanced examples. Most of her examples for finding documents have a "*find* _____" or "*search for* _____". (These types of searches are referred to as *search for* methods.) ELISE also assumes that a good worker takes the initiative. She often automatically searches when a person's statement implies a request for a search. However she sometimes acts cautiously because she knows she has limited understanding of human language. If she cannot determine the intent of a person's statement, she will ask the person to verify that they are asking her to do a search. Sometimes she may even say that she doubts that the search request will yield any good results. ELISE can best be summarized as a search agent who is polite and humble at all times, suggestive whenever she has the opportunity, and cautious whenever she feels like she is being asked to make an off-topic search query. If ELISE truly cannot understand what the user asks of her, she just plays dumb and says she does not understand.

H. Do the task questions encourage the user to refine a search?

There is nothing in the task sheet that explicitly asks the user to refine a search. However since the underlying search engine is made to have limited accuracy, users will most likely need to make some refinements if they want the document to appear as one of the first 3 titles in the results. If users do not submit refinements to the search, then they will most likely spend more time looking through the results of each query.

IV. FINDINGS

1. Time

ANOVA tests were run on the times to complete each question. With the exception of question 3, all times for each question did not reveal any conclusive evidence. The ANOVA tests for question 3 showed that time did have a statistically significant difference. The difference is in favor of the traditional interface, with the effective difference being about 3 minutes. The ANOVA test for the total time to complete the task sheets also showed a statistically significant difference in total time in favor of the traditional interface. The effective difference of the total time averages is about 6 minutes.

Despite the statistically inconclusive evidence for time on each question, the average times for ELISE were higher on each question. The higher averages on question 1, 2, and 4 may account for the other 3 minutes of the difference in total time, but such a claim is admittedly difficult to support. Examining the statistical data for the total time to complete the task sheets show that the mean total time for the traditional interface was 7 minutes and 36 seconds, and the mean total time for ELISE users was 13 minutes and 36 seconds. Since the mean total time for ELISE took 1.78 times longer than the traditional interface and the ANOVA test indicates

statistical significance, then it follows that the traditional interface has an effective advantage over the conversational interface in terms of time. (See [] in APPENDIX I.)

A statistically significant learning effect was noticed in the time data. People who started with the traditional interface took more time with ELISE, except those people did not take as much time using ELISE as the people starting with ELISE did. The people who started with ELISE took less time when using the traditional interface, except those people did took more time using the traditional interface than those who started with the traditional interface. (See [] in the APPENDIX J.)

2. Number of Searches Submitted

The ANOVA test for questions 1 and 2 showed inconclusive evidence for the number of documents opened. However the data for these questions show a very narrow range: one to three searches. For question 1, only one person in group A submitted two searches and three people in group B submitted 2 searches. Such evidence overwhelmingly suggests that if a statistically significant difference does indeed exist, the effective (practical) difference is very marginal.

The time results for question 2 are similar to question 1, with ANOVA indicating inconclusive evidence. However everybody in group A made 1 search, providing no variance. A majority of 5 people in group B made 1 search with a one person submitting 2 searches and another submitting 3 searches. Even if a statistically significant difference exists, it is likely to suggest the traditional interface is marginally better than the conversational one.

The ANOVA results for question 3 show significant performance in favor of the traditional interface. Examining the recorded dialogs for ELISE seems to indicate that people refined their searches. Although no definite trends can be stated, three reasons for refinement seem implicit in the dialogs with ELISE:

- ELISE did not choose the intended words in a search.
- ELISE included possibly unintended words in a search.
- Users experimented with the natural language capabilities of ELISE.

The mean for group A was 1.6 and the mean for group B was 3. It is debatable whether or not this difference in means does provide any practical improvement.

The ANOVA results for question 4 showed a strong indication that neither interface provides any advantage in terms of searches. These results conflict with the results in question 3, which does bring doubt to the timing results for questions 3 and 4.

When examining the total number of searches for each interface, neither interface yielded conclusive evidence. This may be largely due to the conflicting results given by questions 3 and 4 as well as the low variance seen in questions 1 and 2.

3. Number of Documents Opened

Only one person in all 14 people sampled opened more than one document for question 1; that person opened only 2. The ANOVA tests indicated that these findings are not significant, but the lack of variance is so rare that it makes sense to conclude that neither interface gave an effective advantage to question 1. Even if one were to argue that a difference could exist, the effective improvement is likely to be marginal.

The ANOVA results for the second question are unreliable. Many people seemed to disregard the assumption stated in the question. The assumption is that if a document's title implies anything about the desired subject, then the title is considered a valid answer. Some people opened documents and actually read the documents.

The ANOVA results for the third and fourth question are inconclusive. When examining the total number of documents opened for each interface, neither interface yielded any kind of conclusive evidence.

4. Accuracy

ANOVA indicated inconclusive results for questions 1 and 2, although this may be caused by a lack of variance. Only one person failed to get full credit for the accuracy on question 1 because that person didn't open the correct document when using ELISE. Everybody else got the full 4 points for accuracy. On the second question, only one person failed to get full credit, and that person got 2 points. With both of these questions, it seems apparent that neither interface really provides an effective advantage in terms of accuracy.

ANOVA indicated that questions 3 and 4 have inconclusive evidence regarding accuracy. Despite the inconclusive nature of the data for individual questions, ANOVA revealed that the total scores for both interfaces show a lack of difference between total accuracy. (See the ANOVA graph in the APPENDIX K.) The equivalence test on total accuracy showed that paired scores will be within 8% in a 96% statistical confidence margin. Since 8% falls under the traditional 10% margin for equivalence and 8% of 16 points is 1.28 points (which is effectively half the points of one question), the results strongly indicated that the two interfaces are

effectively equivalent in terms of accuracy. (See [] for the data on the equivalence test in APPENDIX L.)

5. User Evaluations

The Wilcoxon Signed Rank tests on the interface evaluation surveys gave no conclusive data regarding users' perceptions of the interfaces' accuracy, speed, usability, and likeableness. Of the 14 people sampled, only 3 people thought that ELISE did not have a believable personality. The Binomial single proportion tests done on the responses for users' preferences indicated the following:

- At a statistical confidence level of 85.4%, people would prefer to use ELISE in workplace scenarios.
- At a statistical confidence level of 93.5%, people would prefer to use ELISE in personal scenarios.

V. DISCUSSION

1. Issues of Time and Accuracy

The immediate implications of this data is that people are more efficient in terms of time when using the traditional interface, but they are just as accurate when they use the conversational ELISE interface. While existing research says that natural language interfaces are just as or more accurate than Boolean interfaces (Turtle, 1996) the issue of time has not been studied in depth. Since the data for number of submitted searches and number of opened documents is unreliable and due to the wide range of possible numbers for both metrics, I cannot make any conclusions on how submitting searches and opening documents contribute to the differences in time.

However some differences in the interface do lend insight into why the traditional interface seems more efficient with time. ELISE gives much more feedback to the user than the traditional interface. She summarizes a search, suggests which documents to examine, and sometimes gives suggestions on how to refine a search. She also conveys this information in full sentences, whereas the traditional interface simply lists the vital statistics of a search in a summary table. Furthermore any examples provided by ELISE have additional dialog. All of ELISE's feedback requires her to print more text than the traditional interface, therefore people have to read more. Evidence for people having read ELISE'S statements is seen on the user evaluations. The people who believe ELISE's personality refer to her dialog when judging her personality. Related to this issue of printed text is the user's contribution to using more text. Examining the log indicates that people who used the traditional interface used about 4 words

per search query. However people who used ELISE reveal that full sentence natural language queries tend to be longer than 4 words. Since it takes more typing to form a natural language query, it follows that it takes more time to type the query.

Another possible reason for the improved time on the traditional interface is that people who did this experiment needed to know how to use a web browser. While nobody had any training on either interface, the traditional interface is similar to the search engines on the Internet. People who have used web browsers are likely to have used the browsers with Internet search engines. So while people have no actual experience with traditional interface used for this experiment, in practice they may have some knowledge that contributed to efficient use.

The ANOVA in the APPENDIX J indicates that people spent more time with ELISE regardless of what order they used the interfaces, adding more support to the hypothesis that people will spend more time to do a task when using a conversational interface. As to the extent of a participant's Internet experience affects the participant's performance with each interface remains unanswered. Further research in this area is needed to make any strong conclusions.

2. Issues of Preference for Personality

The evidence gathered on the user preferences indicate that people prefer to use ELISE, which agrees with the existing positive claims of natural language interfaces and virtual personalities. (Capindale, Crawford, 1990; Lester, Converse, Kahler, Barlow, Stone, and Bhoga, 1997) Only 3 out of the 14 people said ELISE was not believable. One person said he had higher expectations of a natural language interface with a personality, claiming that ELISE did not understand basic sentences. Examining his dialog with ELISE revealed that obstacles did show up in his search.

This person typed queries in the traditional format at first, and ELISE was unable to understand the intent of his statement. ELISE responded by choosing words that were related to Photoshop, and then asking the user if he wanted her to search for those keywords. Eventually this person started typing queries that started with *search for* and found documents successfully. Another person cited that ELISE lacked mood and was too much like a computer. Another person cited that ELISE did not seem much different from other search engines. Analyzing this person's dialog indicated that the person primarily used *search for* queries, explaining why the person would think ELISE is much like traditional search engines. People who did think ELISE was believable cited the following reasons. Incidentally, some of these reasons also appeared as reasons for preferring ELISE:

- When summarizing a search she conveys judgement on the quality of the highest scoring document and the quantity of the search.
- She is very polite and says things similar to that of a librarian.
- She uses complete sentences like a person and understands what I say.
- She smiles and sometimes has attitude.
- It would be fun to talk to ELISE when attempting to avoid work at the workplace.

People's general preference for the conversational interface over the traditional interface has implications for how interfaces might be designed. If one of the goals for the interface is to make its interaction more enjoyable, and if time efficiency is not that important, then it makes sense to use a conversational interface. The possible niche for a conversational search interface may be fore the purposes of teaching, or reducing negative feelings towards such software. Such an idea is consistent with the studies that concluded that a computer with personality has positive effects on learning (Lester, et al, 1997) and that people respond positively to natural language interfaces

(Capindale, et al. 1990). Also if people are looking for an alternative to the traditional search engines on the Internet, a conversational one may be the search engine for them. However if time efficiency is more important than user preference, then a traditional interface should be the choice in such a situation.

3. Issues of Learning

Figure 30 – Generic Learning Curves for the Interfaces

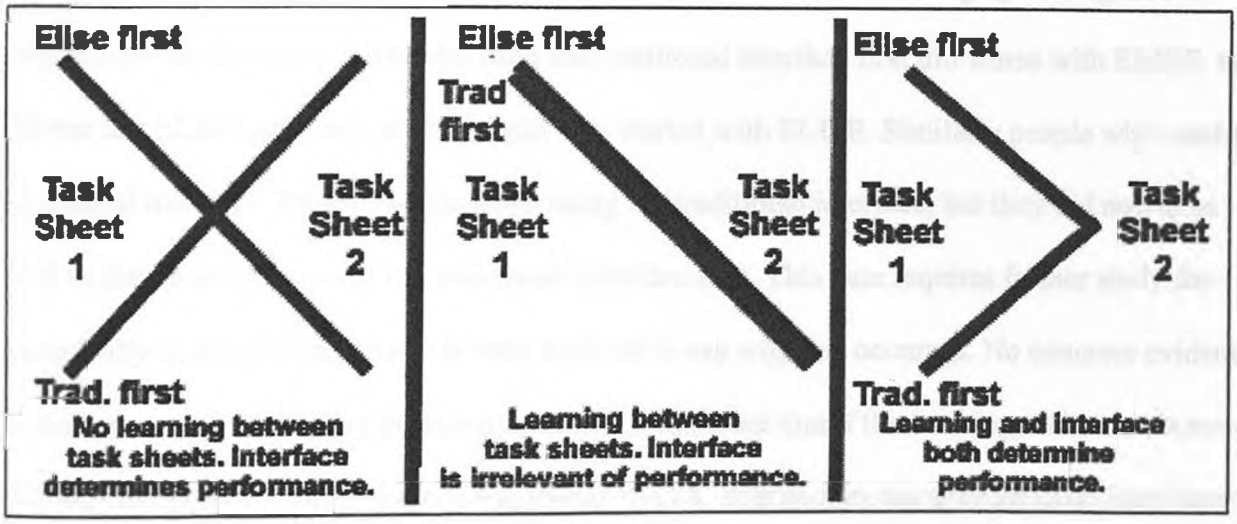


Figure 30 shows three graphs. Each graph represents how time efficiency can change when using the different interfaces sequentially. Light gray lines show how time efficiency can change for those who use the traditional interface first. Dark gray lines show how time efficiency can change for those who use ELISE first. A line that has a negative slope shows that the user improved time-wise with the interface used for Task Sheet 2, whereas a line that has a positive slope shows that the user did worse.

One statistically significant finding that was unexpectedly observed involved the learning curve between interfaces. (See [] in APPENDIX J.) If no learning existed between using task sheet 1 and task sheet 2, then the ANOVA 3 graph would look like the right-most graph pictured

in Figure 30. Such a graph indicates two things. The people who used ELISE first do as well as people who used ELISE second, and the people who used the traditional interface first do as well as those who used the traditional interface second. They learn nothing from the first task sheet. Each interface is always going to perform as best as it can regardless of the order it was used. If learning is everything between the two task sheets, then the ANOVA 3 graph would resemble the middle graph in Figure 30. Such a case would mean that regardless of which interface people used first, they always do better on the second task sheet.

The actual results in the ANOVA 3 graph resemble the left-most graph in Figure 30. While it is true that the people who used the traditional interface first did worse with ELISE, they did not use ELISE as much as the people who started with ELISE. Similarly people who used the traditional interface first did better when using the traditional interface, but they did not do as well as the people who used the traditional interface first. This data requires further study for statistically conclusive results; it is very difficult to say why this occurred. No concrete evidence in the logs suggests reasons for these results. If it turns out that if the learning difference between the two interfaces is valid, then conversational search engines may not provide good interfaces for teaching users about Boolean searches. I must note that this study does not deal with a hybrid interface that uses traditional Boolean queries, but also allows the user to receive feedback and help through a conversational interface. Such an interface may be have the best of both interfaces, although any conclusions would need to come from a separate study.

4. Overall Conclusions

While the existing research examined for this study indicates good practicality for natural language interfaces, the results of my experiment indicate that natural language interface with virtual a personality provides no effective advantage over the traditional interface in a context of search engines. One can argue that this study excludes the most powerful aspects of natural language, such as identifying the context of keywords. Such people can also argue that advanced industry-level Boolean search engines are not as efficient as comparable natural language search engines. In such cases I cannot rightly dispute those points because I make no conclusions against natural language interfaces in general. If the conversational interface in this study used context-sensitive natural language queries, and the traditional interface was an advanced Boolean search engine, it is likely that existing research would prevail – natural language interfaces will be more efficient than the advanced Boolean search engines.

One might conclude that advanced Boolean search engines are inferior to natural language interfaces and then provide the corollary that simple traditional interfaces are inferior to natural language interfaces. This study puts doubt to such a claim. The results show that a conversational interface does not provide any effective advantage over a traditional interface, when both interfaces work with an underlying Boolean search engine. Using the results from this study and from previous studies, the following table (Figure 31) can be constructed to determine which interface would work better given the following circumstances.

Figure 31 – Table for Determining Which Interface to Use Given Certain Conditions

	Simple Query Grammar	Complex Query Grammar
Time efficiency matters more than user preference	*Traditional*	*Natural Language*
User satisfaction matters more than time preference	*Conversational*	*Natural Language*

The columns represent the choice between an underlying search engine having a *simple* or *complex* query grammar. A definition of a simple query grammar would be *any query grammar that only uses keywords and the basic Boolean operators of AND, NOT, and OR.* Complex grammars are non-simple grammars. Such grammars might allow people to specify how words in the query are related to each other. The two rows of the table determine whether or not time efficiency matters more than user preference.

Natural language interfaces are preferable when working with a complex query as indicated by the research done by Turtle in 1996 and implied by other studies like Capindale, et al. At the same time if the underlying search engine uses a simple query grammar and the project goals favor time efficiency over user satisfaction, a traditional interface may be the choice interface for such an endeavor.

I should restate that this study evaluates a natural language interface that *has a personality* against an interface inspired from existing Internet search engines, whereas one could evaluate a natural language interface without a personality against a traditional interface. Although research on search engines shows positive results for natural language interfaces, natural language does not necessarily imply the presence of a personality or conversational ability. One can conclude that a conversational interface will do better than a traditional interface

if a complex query grammar is involved. This is because a good conversational interface would have all of the properties that make a natural language interface more successful than a traditional one. However to conclude that a conversational interface will do better than a plain natural language interface in such scenarios may be erroneous because the two interfaces have yet to be compared against each other. Even though those two interfaces have not been compared it seems possible that a conversational interface may have worse time efficiency than a plain natural language interface. The reason for such a claim follows from the same reasons the traditional interface was more time efficient than the conversational interface. The more text the computer prints, the more people have to read.

While this study does not attempt to examine how well a conversational interface teaches a person about doing searches, a statistically unexpected result indicated there is a learning effect according to which interface was used first. The learning effect goes against the conversational interface because people who used the traditional interface the second time took more time than the people who used the traditional interface first. Even taking learning into account shows that people are faster with the traditional interface against conversational interfaces.

Despite the faster performance when using the traditional interface there is some statistical implication that people would prefer to use a conversational interface over a traditional interface. If the statistical implications in this study are true, they simply agree with existing research that people prefer natural language. Since some of the reasons people provided for preference were related to the conversational aspects, then it is somewhat implied that people prefer personality. I cannot say whether the preference for natural language and virtual personalities are inherent to people. That field of study involves more psychology, and I am not a reliable critic for making conclusions about the psychology of socially active humans. However

this research casts doubt on whether or not our technologically oriented society really needs a talking computer. It may be fun and nice to have a talking computer, but it might be just more time efficient simply to have a functional computer.

APPENDIX A

BLANK ANSWER SHEET

Unique ID	Task Sheet

Question 1

Circle One YES NO

Question 2

Document title:_____

Question 3

Maximum Number:_____

Document title:_____

Question 4

Document title:_____

APPENDIX B

SCREENING QUESTIONNAIRE

How often do you usually use a computer during the week?
One day a week or less – 0 points
2 or 3 days per week – 3 points
4 or 5 days per week – 6 points
Over 5 days per week – 9 points

On a normal day when you use a computer, how many hours do you spend using one that day?
One hour or less per day – 0 points
2 or 3 hours per day – 3 points
4 or 5 hours per day – 6 points
Over 5 hours per day – 9 points

How often do you normally use the following software in a week?
Web browsers such as Internet Explorer and Netscape.

One hour or less – 0 points
1 - 3 hours – 2 points
3 - 5 hours – 4 points
Over 5 hours – 6 points

Other Internet & Office Programs such as email, Internet chat, word-processing, ICQ, spreadsheets, databases, presentation software.

One hour or less – 0 points
1 - 3 hours – 1 points
3 - 5 hours – 2 points
Over 5 hours – 3 points

Other graphics programs such as Adobe Illustrator, Macromedia Freehand, Paint Shop Pro, 3D Studio, SoftImage.

One hour or less – 0 points
1 - 3 hours – 2 points
3 - 5 hours – 3 points
Over 5 hours – 4 points

Give your best estimate of how well you perform searches using Internet search engines.

It usually takes no more than 5 minutes to find what I want. I find what I want within the first 5 attempts. I feel like I find things efficiently on the world-wide-web.	6 points
It usually takes no more than 10 minutes to find what I want. I find what I want around 5-10 attempts and refinements. I feel like I find things alright on the world-wide-web.	4 points
It usually takes me more than 10 minutes to find what I want. I find what I want after 10 attempts and refinements. I feel like it takes me forever to find what I want on the world wide web.	2 points
I don't know. I don't do enough Internet searches to really know how long it takes.	0 points

How much do you know about Boolean algebra?

I know Boolean algebra well and I can understand a phrase like (True AND False) OR ((True AND True) AND (True OR False)) without too much thinking.	9 points
I know what it would mean if somebody said to look for: (cats AND dogs) OR (cats AND fish) but NOT (dogs AND fish)	6 points
I know the difference between looking for (cats AND dogs) and looking for (cats OR fish).	3 points
I have no idea how to interpret any of the phrases above, or how I would apply such information when doing a search.	0 points

APPENDIX C

TASK SHEET 1

You are doing the first task sheet
in a series of two task sheets.
This task sheet has 4 questions.
You will not be penalized in any way
for incorrect or incomplete answers
so take your time and answer well.
Please note that you cannot
go back to previous questions after
you have proceeded from them.
Click here to start

Task Sheet 1

Question 2 of 4

List the title of one document
that explains how to copy a layer.
You do not need to read any documents.
If a document's title has anything
to do with how to copy a layer
then it's an acceptable answer.

Remember you cannot go back to this question
once you go on to the next one.
Click here for the next question.

Task Sheet 1

Question 3 of 4

The layers palette has a maximum number of active layers.

What is that maximum?

List the number and the document title where you got that information.

Remember you cannot go back to this question once you go on to the next one.

Click here for the next question.

Task Sheet 1

Question 4 of 4

Find a document that shows step by step how to create some text that appears to be on fire. (see example below).

Remember you cannot go back after you finish.

Click here to finish.

You have finished Task Sheet 1.
Please evaluate the software you just used
by filling out the user satisfaction survey.
You can find the user satisfaction survey on
the last page of your answer booklet.
After that you will go on to do Task Sheet 2
and fill out 2 more surveys.

APPENDIX D

TASK SHEET 2

You are doing the second task sheet
in a series of two task sheets.
This task sheet has 4 questions.
You will not be penalized in any way
for incorrect or incomplete answers
so take your time and answer well.
Please note that you cannot
go back to previous questions after
you have proceeded from them.
Click here to start

Task Sheet 2

Question 1 of 4

Find the document titled "Printer resolution".

Is there a graphic (picture) in this document?

Remember you cannot go back to this question

once you go on to the next one.ts2q2.html

Click here for the next question.

Task Sheet 2

Question 2 of 4

List the title of one document that explains how to export an image.

You do not need to read any documents.

If a document's title has anything to do with how to export an image of any kind then it's an acceptable answer.

Remember you cannot go back to this question once you go on to the next one.

Click here for the next question.

Task Sheet 2

Question 3 of 4

An indexed-color palette has a maximum number of colors it can contain.

What is that maximum?

List the number and the document title where you got that information.

Remember you cannot go back to this question once you go on to the next one.

Click here for the next question.

Task Sheet 2

Question 4 of 4

Find a document that explains step by step how to add a shadow to a picture of a standing person.

(see example below)

Remember you cannot go back after you finish.

Click here to finish.

You have finished Task Sheet 2.
Please evaluate the software you just used
by filling out the user satisfaction survey.
You can find the user satisfaction survey on
the last page of your answer booklet.
In addition to the user-satisfaction survey please
fill out the software comparison questionnaire.
The questionnaire follows the
user-satisfaction survey.

APPENDIX E

TRADITIONAL INTERFACE EVALUATION SURVEY

Office Use Only: ID:_____

1. Did the search agent work well with providing accurate and relevant material after doing its searches? Circle One.

- Very much so – I found what I needed to find without too much trouble.
- Sometimes – It had some trouble working with the things I typed from time to time.
- Not often – I had a lot of trouble getting it to find things properly.

2. Did the search agent give results in a timely manner? Circle One.

- Very much so – It responded instantaneously.
- Somewhat – Every now and then it seemed slow to me.
- Not at all – It was slow.

3. Was the search agent's user interface easy to understand and use? Circle One.

- Yes – I understood how to do searches almost immediately.
- Somewhat – It was awkward at first but I figured it out eventually.
- No – I was confused many times and I wasn't sure how I was supposed to do a search.

4. Did you like how the interaction occured with the search agent? (ie, Did you enjoy using the software in any way?) Why or why not?

APPENDIX F

ELISE INTERFACE EVALUATION SURVEY

Office Use Only: ID:_____

1. Did Elise work well with providing accurate and relevant material after doing its searches? Circle One.
   - Very much so – I found what I needed to find without too much trouble.
   - Sometimes – It had some trouble working with the things I typed from time to time.
   - Not often – I had a lot of trouble getting it to find things properly.
2. Did Elise give results in a timely manner? Circle One.
   - Very much so – It responded instantaneously.
   - Somewhat – Every now and then it seemed slow to me.
   - Not at all – It was slow.
3. Was the Elise's user interface easy to understand and use? Circle One.
   - Yes – I understood how to do searches almost immediately.
   - Somewhat – It was awkward at first but I figured it out eventually.
   - No – I was confused many times and I wasn't sure how I was supposed to do a search.

4. Do you think Elise has a believable personality? In other words, do you believe that Elise makes a successful attempt at trying to seem like a human? This question is **not asking** whether you think the computer is indistinguishable from a human. It is understood that no computer has yet been able to pass as a human in conversation. Answer yes or no and list reasons if you possible.

5. Did you like how the interaction occured with the search agent? (ie, Did you enjoy using the software in any way?) Why or why not?

APPENDIX G

INTERFACE COMPARISON SURVEY

Office Use Only: NO ID:_____

1. In comparing the two interfaces, which do you believe gave more accurate and relevant results? Circle one. Circle both if both are equal.

- No conversation, user driven interface
- Conversational Elise interface

2. In comparing the two interfaces, which do you believe gave results in a more timely fashion? Circle One. Circle both if both are equal.

- No conversation, user driven interface
- Conversational Elise interface

3. In comparing the two interfaces, which do you believe was easier to use. Circle One. Circle both if both are equal.

- No conversation, user driven interface
- Conversational Elise interface

4. Of the two types of search engines, which would you prefer to use in a workplace environment and why?

5. Of the two types of search engines, which would you prefer to use in a personal situation (ie, non-professional, casual, relaxed environments) and why?

APPENDIX H

EXPLANATION OF 1-2*(1-P)

The binomial distribution (P) is the probability that we would get *x* number of people who would choose ELISE over the traditional in a sample of *n* people. 1-P is the probability that we would get more than *x* successes from *y* people. 2*(1-P) is the probability that the people would pick ELISE or the traditional. 1-2*(1-P) is the probability that we would get up to that people who would choose ELISE or a comparable number of people who would choose the traditional.

APPENDIX I

ANALYSIS OF VARIANCE OF TOTAL TIME BY INTERFACE

TABLE 1. Analysis of Variance of Total Time by Interface

ANOVA Table for Interface

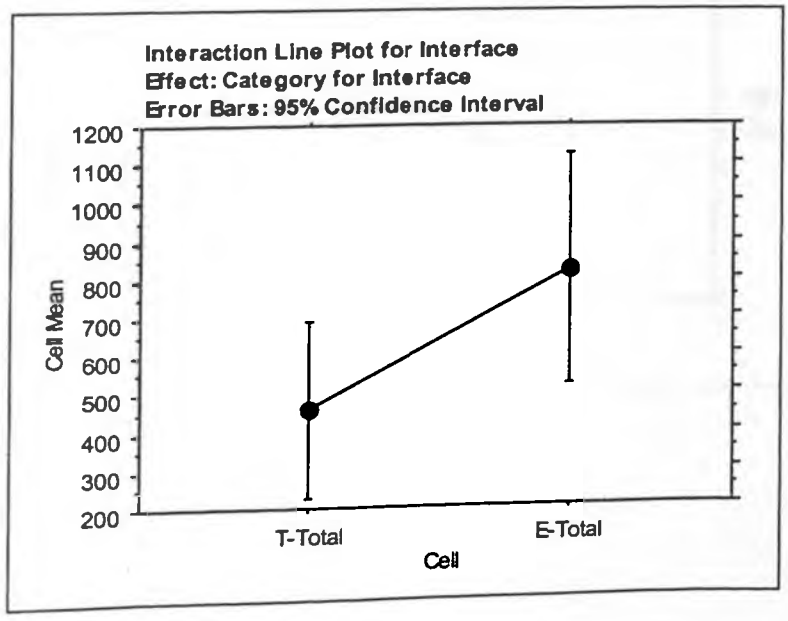
	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	13	3923029.429	301771.495				
Category for Interface	1	918032.143	918032.143	6.866	.0212	6.866	.681
Category for Interface * Subject	13	1738091.857	133699.374				

TABLE 2. Means of Total Time by Interface

Means Table for Interface
Effect: Category for Interface

	Count	Mean	Std. Dev.	Std. Err.
T-Total	14	458.071	398.331	106.458
E-Total	14	820.214	526.121	140.612

GRAPH 1. Analysis of Variance Showing the Mean Total Times for the Traditional Interface (T-Total) and the ELISE Interface (E-Total)



APPENDIX J

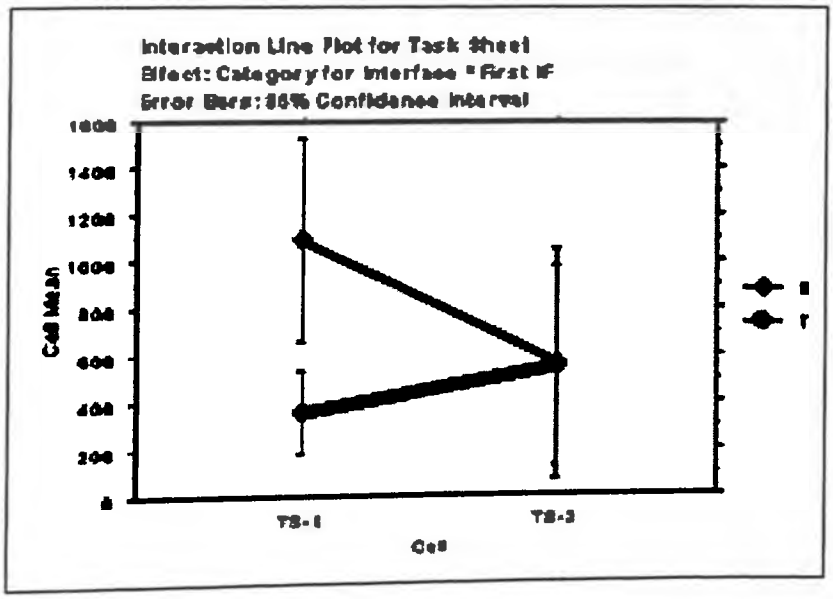
ANALYSIS OF VARIANCE: INTERACTION PLOT OF TOTAL TIME TAKING FIRST INTERFACE INTO ACCOUNT

TABLE 3. Analysis of Variance of Total Time by Interface and by Order of Interface

ANOVA Table for Task Sheet

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
First IF	1	976142.286	976142.286	3.975	.0694	3.975	.439
Subject(Group)	12	2946887.143	245573.929				
Category for Interface	1	201281.286	201281.286	1.572	.2338	1.572	.202
Category for Interface * First IF	1	918032.143	918032.143	7.168	.0201	7.168	.694
Category for Interface * Subject(Group)	12	1536810.571	128067.548				

GRAPH 2. Analysis of Variance of Total Time by Interface and by Order of Interface



APPENDIX K

ANALYSIS OF VARIANCE OF TOTAL ACCURACY BY INTERFACE

TABLE 4. Analysis of Variance of Total Accuracy by Interface

ANOVA Table for Interface

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	13	129.750	9.981				
Category for Interface	1	.036	.036	.016	.9020	.016	.052
Category for Interface * Subject	13	29.464	2.266				

TABLE 5. Means of Total Accuracy by Interface

Means Table for Interface
Effect: Category for Interface

	Count	Mean	Std. Dev.	Std. Err.
TS-1	14	14.286	2.335	.624
TS-2	14	14.214	2.607	.697

GRAPH 3. Analysis of Variance Showing the Mean Total Accuracy for the Traditional Interface (T-Total) and the ELISE Interface (E-Total)

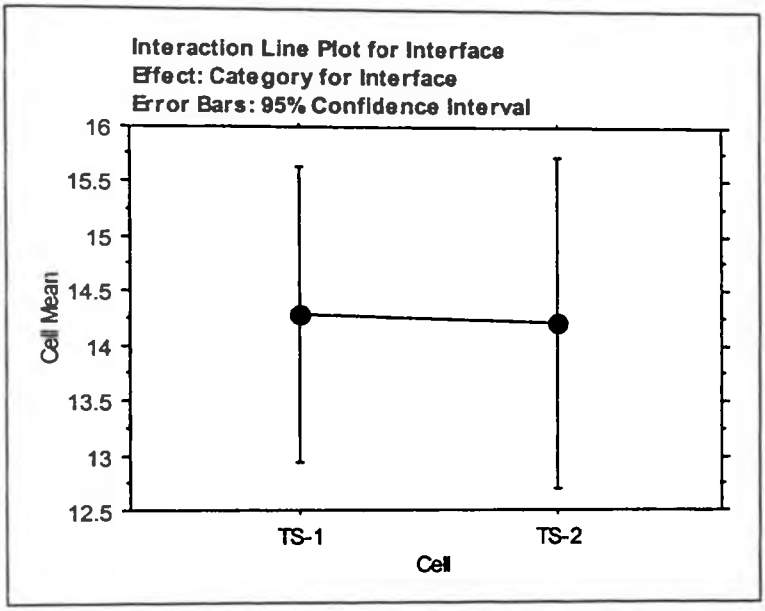


TABLE 6. Equivalency Test: Total Accuracy

Total possible	16
Delta percentage	0.08
Delta	1.28
Mean Difference	0.071428571
SE	0.569019648
T-value	0.125529183
P-value	0.038928477
Comment	For 8% difference in Total Accuracy

APPENDIX L

RAW DATA

PLEASE NOTE THAT RAW DATA IS NOT MENTIONED IN THE TEXT BUT DOES ALLOW THE CURIOUS AND READER TO CHECK THE PREVIOUSLY SHOWN GRAPHS AND TABLES.

TABLE 7. Time to Complete Each Question by Task Sheet.

ID	TS1Q1	TS1Q2	TS1Q3	TS1Q4	TS2Q1	TS2Q2	TS2Q3	TS2Q4	TOTAL1	TOTAL2	AVE1	AVE2
07011230	186	179	123	37	165	124	910	238	525	1437	131.25	359.25
07081000	120	257	432	400	101	125	489	555	1209	1270	302.25	317.5
07151600	145	167	643	324	75	103	487	717	1279	1382	319.75	345.5
07181900	104	216	187	634	35	49	78	115	1141	277	285.25	69.25
07181515	53	50	205	301	30	52	359	259	609	700	152.25	175
07201430	105	80	1177	441	40	81	103	16	1803	240	450.75	60
07211130	72	174	95	155	86	45	394	190	496	715	124	178.75
07221130	66	115	444	606	55	46	55	33	1231	189	307.75	47.25
07231430	115	93	50	28	60	70	202	48	286	380	71.5	95
07291400	34	33	95	19	59	25	181	29	181	294	45.25	73.5
07291530	28	17	20	51	12	26	8	12	116	58	29	14.5
07301300-	63	68	189	223	50	55	160	175	543	440	135.75	110
08011300	32	68	79	257	18	15	41	48	436	122	109	30.5
08011700	29	58	23	170	44	48	62	103	280	257	70	64.25
AVERAGE	82.28571	112.5	268.7143	260.4286	59.28571	61.71429	252.0714	181.2857	723.9286	554.3571	180.9821	138.5893

TABLE 8. Time to Complete Each Question by Interface.

ID	T-Q1	T-Q2	T-Q3	T-Q4	E-Q1	E-Q2	E-Q3	E-Q4	T-Total	E-Total	T-Ave	E-Ave
07011230	186	179	123	37	165	124	910	238	525	1437	131.25	359.25
07081000	101	125	489	555	120	257	432	400	1270	1209	317.5	302.25
07151600	75	103	487	717	145	167	643	324	1382	1279	345.5	319.75
07181900	35	49	78	115	104	216	187	634	277	1141	69.25	285.25
07181515	53	50	205	301	30	52	359	259	609	700	152.25	175
07201430	40	81	103	16	105	80	1177	441	240	1803	60	450.75
07211130	72	174	95	155	86	45	394	190	496	715	124	178.75
07221130	55	46	55	33	66	115	444	606	189	1231	47.25	307.75
07231430	115	93	50	28	60	70	202	48	286	380	71.5	95
07291400	34	33	95	19	59	25	181	29	181	294	45.25	73.5
07291530	28	17	20	51	12	26	8	12	116	58	29	14.5
07301300-	50	55	160	175	63	68	189	223	440	543	110	135.75
08011300	18	15	41	48	32	68	79	257	122	436	30.5	109
08011700	29	58	23	170	44	48	62	103	280	257	70	64.25
AVERAGE	63.64286	77	144.5714	172.8571	77.92857	97.21429	376.2143	268.8571	458.0714	820.2143	114.5179	205.0536

TABLE 9. Number of Searches Submitted for Each Question by Task Sheet.

ID	TS1Q1	TS1Q2	TS1Q3	TS1Q4	TS2Q1	TS2Q2	TS2Q3	TS2Q4	TOTAL1	TOTAL2	AVE1	AVE2
07011230	1	1	1	1	1	1	5	2	4	9	1	2.25
07081000	1	1	2	2	1	1	2	5	6	9	1.5	2.25
07151600	1	1	4	3	1	1	3	4	9	9	2.25	2.25
07181900	2	3	6	3	1	1	2	2	14	6	3.5	1.5
07181515	2	1	2	5	1	1	5	4	10	11	2.5	2.75
07201430	2	1	1	6	1	1	1	1	10	4	2.5	1
07211130	1	1	2	6	1	1	3	1	10	6	2.5	1.5
07221130	1	1	2	4	1	1	1	1	8	4	2	1
07231430	1	1	1	1	1	1	3	1	4	6	1	1.5
07291400	1	1	2	1	2	1	6	2	5	11	1.25	2.75
07291530	1	1	1	2	1	2	1	1	5	5	1.25	1.25
07301300-	1	1	2	3	1	1	2	2	7	6	1.75	1.5
08011300	1	1	1	3	1	1	2	2	6	6	1.5	1.5
08011700	1	1	1	5	1	1	1	2	8	5	2	1.25
AVERAGE	1.214286	1.142857	2	3.214286	1.071429	1.071429	2.642857	2.142857	7.571429	6.928571	1.892857	1.732143

TABLE 10. Number of Searches Submitted for Each Question by Interface.

ID	T-Q1	T-Q2	T-Q3	T-Q4	E-Q1	E-Q2	E-Q3	E-Q4	T-Total	E-Total	T-Ave	E-Ave
07011230	1	1	1	1	1	1	5	2	4	9	1	2.25
07081000	1	1	2	5	1	1	2	2	9	6	2.25	1.5
07151600	1	1	3	4	1	1	4	3	9	9	2.25	2.25
07181900	1	1	2	2	2	3	6	3	6	14	1.5	3.5
07181515	2	1	2	5	1	1	5	4	10	11	2.5	2.75
07201430	1	1	1	1	2	1	1	6	4	10	1	2.5
07211130	1	1	2	6	1	1	3	1	10	6	2.5	1.5
07221130	1	1	1	1	1	1	2	4	4	8	1	2
07231430	1	1	1	1	1	1	3	1	4	6	1	1.5
07291400	1	1	2	1	2	1	6	2	5	11	1.25	2.75
07291530	1	1	1	2	1	2	1	1	5	5	1.25	1.25
07301300-	1	1	2	2	1	1	2	3	6	7	1.5	1.75
08011300	1	1	2	2	1	1	1	3	6	6	1.5	1.5
08011700	1	1	1	5	1	1	1	2	8	5	2	1.25
AVERAGE	1.071429	1	1.642857	2.714286	1.214286	1.214286	3	2.642857	6.428571	8.071429	1.607143	2.017857

TABLE 11. Number of Documents Opened for Each Question by Interface.

ID	TS1Q1	TS1Q2	TS1Q3	TS1Q4	TS2Q1	TS2Q2	TS2Q3	TS2Q4	TOTAL1	TOTAL2	AVE1	AVE2
07011230	1	1	2	1	1	0	5	1	5	7	1.25	1.75
07081000	1	1	2	8	1	0	5	1	12	7	3	1.75
07151600	1	1	4	9	1	1	4	9	15	15	3.75	3.75
07181900	1	1	4	8	1	0	6	12	14	19	3.5	4.75
07181515	1	1	1	8	1	2	2	5	11	10	2.75	2.5
07201430	1	1	4	10	1	0	8	5	16	14	4	3.5
07211130	2	1	1	15	1	2	5	1	19	9	4.75	2.25
07221130	1	2	4	4	1	1	7	4	11	13	2.75	3.25
07231430	1	1	6	12	1	0	4	1	20	6	5	1.5
07291400	1	0	1	1	1	1	3	1	3	6	0.75	1.5
07291530	1	1	3	1	1	0	7	2	6	10	1.5	2.5
07301300-	1	1	2	4	1	0	3	3	8	7	2	1.75
08011300	1	1	2	4	1	0	4	3	8	8	2	2
08011700	1	3	2	7	1	1	2	3	13	7	3.25	1.75
AVERAGE	1.071429	1.142857	2.714286	6.571429	1	0.571429	4.642857	3.642857	11.5	9.857143	2.875	2.464286

TABLE 12. Number of Documents Opened for Each Question by Task Sheet.

ID	T-Q1	T-Q2	T-Q3	T-Q4	E-Q1	E-Q2	E-Q3	E-Q4	T-Total	E-Total	T-Ave	E-Ave
07011230	1	1	2	1	1	0	5	1	5	7	1.25	1.75
07081000	1	0	5	1	1	1	2	8	7	12	1.75	3
07151600	1	1	4	9	1	1	4	9	15	15	3.75	3.75
07181900	1	0	6	12	1	1	4	8	19	14	4.75	3.5
07181515	1	1	1	8	1	2	2	5	11	10	2.75	2.5
07201430	1	0	8	5	1	1	4	10	14	16	3.5	4
07211130	2	1	1	15	1	2	5	1	19	9	4.75	2.25
07221130	1	1	7	4	1	2	4	4	13	11	3.25	2.75
07231430	1	1	6	12	1	0	4	1	20	6	5	1.5
07291400	1	0	1	1	1	1	3	1	3	6	0.75	1.5
07291530	1	1	3	1	1	0	7	2	6	10	1.5	2.5
07301300-	1	0	3	3	1	1	2	4	7	8	1.75	2
08011300	1	0	4	3	1	1	2	4	8	8	2	2
08011700	1	3	2	7	1	1	2	3	13	7	3.25	1.75
AVERAGE	1.071429	0.714286	3.785714	5.857143	1	1	3.571429	4.357143	11.42857	9.928571	2.857143	2.482143

TABLE 13. Total Accuracy for Each Question by Task Sheet.

ID	TS1Q1	TS1Q2	TS1Q3	TS1Q4	TS2Q1	TS2Q2	TS2Q3	TS2Q4	TOTAL1	TOTAL2	AVE1	AVE2
07011230	50	50	50	50	50	50	15	50	200	165	50	41.25
07081000	0	50	48	0	50	50	0	0	98	100	24.5	25
07151600	50	50	50	50	50	50	50	50	200	200	50	50
07181900	50	50	50	50	50	50	50	50	200	200	50	50
07181515	50	50	15	50	50	50	50	50	165	200	41.25	50
07201430	50	50	48	30	50	50	50	50	178	200	44.5	50
07211130	50	50	48	30	50	50	50	30	178	180	44.5	45
07221130	50	50	48	0	50	50	20	40	148	160	37	40
07231430	50	50	48	50	50	50	50	50	198	200	49.5	50
07291400	50	50	20	50	50	50	50	50	170	200	42.5	50
07291530	50	25	20	50	50	50	50	50	145	200	36.25	50
07301300-	50	50	50	50	50	50	50	50	200	200	50	50
08011300	50	50	48	50	50	50	50	50	198	200	49.5	50
08011700	50	50	50	50	50	50	50	20	200	170	50	42.5
AVERAGE	46.42857	48.21429	42.35714	40	50	50	41.78571	42.14286	177	183.9286	44.25	45.98214

TABLE 14. Total Accuracy for Each Question by Interface.

ID	T-Q1	T-Q2	T-Q3	T-Q4	E-Q1	E-Q2	E-Q3	E-Q4	T-Total	E-Total	T-Ave	E-Ave
07011230	50	50	50	50	50	50	15	50	200	165	50	41.25
07081000	50	50	0	0	0	50	50	0	100	100	25	25
07151600	50	50	50	50	50	50	50	50	200	200	50	50
07181900	50	50	50	50	50	50	50	50	200	200	50	50
07181515	50	50	15	50	50	50	50	50	165	200	41.25	50
07201430	50	50	50	50	50	50	50	30	200	180	50	45
07211130	50	50	50	30	50	50	50	30	180	180	45	45
07221130	50	50	20	40	50	50	50	0	160	150	40	37.5
07231430	50	50	50	50	50	50	50	50	200	200	50	50
07291400	50	50	25	50	50	50	50	50	175	200	43.75	50
07291530	50	25	25	50	50	50	50	50	150	200	37.5	50
07301300-	50	50	50	50	50	50	50	50	200	200	50	50
08011300	50	50	50	50	50	50	50	50	200	200	50	50
08011700	50	50	50	50	50	50	50	20	200	170	50	42.5
AVERAGE	50	48.21429	38.21429	44.28571	46.42857	50	47.5	37.85714	180.7143	181.7857	45.17857	45.44643

TABLE 15. User Evaluation Results for Interface Evaluations

Question 1: Accuracy/Relevancy
Question 2: Speed
Question 3: Usability
Question 4: Likeability

T: Traditional Interface Evaluation
E: ELISE Interface Evaluation
BELIEVE: Did the user believe ELISE's personality?

ID	T1	T2	T3	T4	E1	E2	E3	E4	BELIEVE
07011230	0	1	0	1	0	1	-1	-1	1
07081000	-1	1	0	0	0	1	0	1	1
07151600	0	1	1	0	0	1	0	0	1
07181900	1	1	1	-1	1	1	1	1	1
07181515	0	1	1	-1	0	1	1	0	0
07201430	1	1	1	0	0	1	1	0	1
07211130	0	1	1	1	1	1	1	1	1
07221130	1	1	1	1	-1	0	1	0	0
07231430	1	1	1	0	1	1	0	1	0
07291400	1	1	0	-1	0	1	0	1	1
07291530	-1	0	-1	-1	0	0	0	1	1
07301300-	0	0	1	1	0	0	1	1	1
08011300	1	1	1	0	1	1	1	1	1
08011700	1	1	1	1	1	0	0	0	1
Totals	**5**	**12**	**9**	**1**	**4**	**10**	**6**	**7**	**11**
Averages	**0.357143**	**0.857143**	**0.642857**	**0.071429**	**0.285714**	**0.714286**	**0.428571**	**0.5**	**0.785714**

TABLE 16. User Evaluation Results for Interface Comparison Survey

Question 1: Accuracy/Relevancy
Question 2: Speed
Question 3: Usability
Question 4: Preference in Workplace
Question 5: Preference in non-Workplace

T indicates user chose the Traditional Interface
E indicates user chose ELISE Interface
B indicates that user said both were equal

ID	Q1	Q2	Q3	Q4	Q5
07011230	T	B	T	T	B
07081000	T	T	E	E	E
07151600	B	B	E	B	B
07181900	B	T	B	E	E
07181515	E	E	E	E	E
07201430	B	B	B	B	B
07211130	E	E	B	E	E
07221130	T	T	T	T	T
07231430	E	B	E	E	E
07291400	T	T	T	T	T
07291530	E	E	E	E	E
07301300-	B	B	E	E	E
08011300	B	B	E	E	E
08011700	T	T	T	T	T

ANNOTATED BIBLIOGRAPHY

Alan Burton, Anthony P. Steward. Effects of Linguistic Sophistication on the Usability of a Natural Language Interface. In *Interacting with Computers*. V.5 n.1 pages 31-59. Butterworth-Heinemann Ltd. 1993.
Notes: I referenced the conclusions regarding linguistic sophistication in a natural language interface. The conclusions in this article imply that enhancements to natural language interfaces do not necessarily improve the interface.

Ruth A. Capindale, Robert G. Crawford. Using a Natural Language Interface with Casual Users. In *International Journal of Man-Machine Studies*. v.32 n.3 pages 341-361. Academic Press. 1990.
Notes: I referenced this article for a positive claim about basic natural language technology involving information retrieval with a database.

T. Erickson. Designing Agents as if People Mattered. In *Software Agents*, pages 79-96. MIT Press, 1997.
Notes: Discusses agents through the design issues and problems of an agent's metaphor and its underlying adaptive purpose. I designed the conversational agent with his idea that an agent needs mechanisms for the user to explicitly determine the details of an agent's task in case the user runs into problems.

J.C. Lester, S.A. Converse, S.E. Kahler, S.T. Barlow, B.A. Stone, R.S. Bhoga, The Persona Effect: Affective Impact of Animated Pedagogical Agents. In *CHI '97 Conference Proceedings on Human Factors in Computing Systems*, pages 359-366. Association for Computing Machinery, 1997.
Notes: Presents the idea that a lifelike character in an interactive learning environment can have a positive effect on a student's learning experience. The study's results also show how different personality types in agents can induce different reactions in students.

C. Nass, J. Steuer, E. Tauber, and H. Reeder. Anthropomorphism, Agency, & Ethopoeia: Computers as Social Actors. In *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems -- Adjunct Proceedings*, pages 111-112. Association for Computing Machinery. 1993.
Notes: Study concluding that minimal social cues can induce computer-literate individuals to use social rules when interacting with agents.

C. Nass, Y. Moon, B.J. Fogg, B. Reeves, and C. Dryer. Can Computer Personalities Be Human Personalities? In *CHI '95. Conference companion on Human factors in Computing Systems*, pages 228-229. Association for Computing Machinery. 1995.
Notes: Study supporting that computer personalities can be created through a minimum set of social cues and that people will respond to those cues in the same way they would to a human.

J. Sidhu, C. J. Hinde. An Analysis of the Use of Natural Language Processing Systems in Business. Three Surveys and a Framework. In *Behaviour and Information Technology*. v.16 n.3 pages.145-157. Taylor & Francis. 1997.
Notes: I cite this article to support the idea that natural language interfaces are not uniformly appropriate in all cases. Successful cases seem to be tied to good planning at particular stages of system development.

Tomek Strzalkowski; Jose Perez-Carballo; Marinescu Mihnea. Natural language information retrieval in digital libraries. In *DL '96. Proceedings of the 1st ACM international conference on Digital libraries.* pages 117-125. Association for Computing Machinery. 1996.
Notes: I used this article for a supporting claim regarding the viability of natural language in real-life information retrieval scenarios.

W.S. N. Reilly. A Methodology for Building Believable Social Agents. In *AGENTS '97. Proceedings of the first international conference on Autonomous agents*, pages 114-121. Association for Computing Machinery, 1997.
Notes: Provides a methodology for building believable social behaviors on a character-by-character basis. I used his idea of using a small set of characteristics to create a personality for my conversational agent.

Martin D. Ringle, Richard Halstead-Nussloch. Shaping User Input: A Strategy for Natural Language Dialogue Design. In *Interacting with Computers*. v.1 n.3 pages 227-244. Butterworth-Heinemann Ltd. 1989.
Notes: I referenced this article for its conclusion regarding unrestricted dialogue in NLIs. I also applied the conclusion in terms of the design of the natural language interface for the search agent.

James L. Rogers, Kenneth I. Howard, John T. Vessey. Using Significance Tests to Evaluate Equivalence Between Two Experimental Groups. In *Psychological Bulletin*. V.113 n.3 pages 533-565. American Psychological Association. 1993.

Howard Turtle. Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In *SIGIR '94. Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval.* pages 212-220. Association for Computing Machinery. 1996.
Notes: I referenced this article for a positive claim about natural language technology involving information retrieval with large databases.