Natural Language Query vs. Keyword Search: Effects of Task Complexity on Search Performance, Participant Perceptions, and Preferences

QianYing Wang, Clifford Nass, and Jiang Hu

Department of Communication, Stanford University,
Stanford, California, USA
{wangqy, nass, huj}@stanford.edu

Abstract. A 2x2 mixed design experiment (*N*=52) was conducted to examine the effects of search interface and task complexity on participants' information-seeking performance and affective experience. Keyword vs. natural language search was the within-participants factor; simple vs. complex tasks was the between-participants factor. There were cross-over interactions such that complex-task participants were more successful and thought the tasks were less difficult and reported more enjoyment and confidence when they used keyword search vs. natural language queries, while the opposite was found for simple-task participants. The findings suggest that natural language search is not the panacea for all information retrieval tasks: task complexity is a critical mediator. Implications for interface design and directions for future research are discussed.

1 Introduction

From punch cards to keyboards to graphical participant interfaces (GUIs) to voice participant interfaces (VUIs), interfaces have evolved to allow increasingly intuitive and natural interactions between participants and computers. Among all the breakthroughs and improvements, the use of natural language (NL) as a means of input and output during human-computer interaction (HCI) is one of the most-researched areas.

One of the potential participant benefits afforded by NL technologies is the reduction in the need for learning and training. The promising future of NL-based conversational interfaces (especially with the presence of computer agents) has been widely lauded by visionaries such as Brenda Laurel [1]. Although no one has yet to be able to claim complete success in natural language generation and processing, progress is continually being made. From text-based software agent (e.g., Microsoft™ Clippy) to speech-recognition customer services automation (e.g., United Airlines' flight information hotline), NL-based technologies have advanced into many areas of daily life.

With the explosion of information brought by the Internet and computers in different forms and sizes, information retrieval has become an integral part of modern life in the information age, demanding participant-friendly and efficient interfaces. In order to provide better search experience for average participants, researchers have applied natural langauge processing (NLP) technology to building information retrieval systems [2-6]. However, usability studies concerning seeking and interacting with information have focused on keyword search (KW) rather than natural language [7-16].

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585 , pp. 106 – 116, 2005.

In this paper, we present a laboratory experiment to examine and compare the usability of two kinds of search interfaces: natural language and keyword search. The study explores how performance limitations of NLP affect participants' perceptions of and preferences for search interfaces. We are especially interested in how *task complexity* affects participants' interaction with information retrieval interfaces.

In the following sections we lay out our experimental design, describe our measured results, offer a discussion of findings, and conclude with design implications.

2 Method

2.1 The Two Interfaces

To ensure that participants would be performing a well-established and typical information acquisition activity, we decided to study interfaces used to obtain frequently-asked information on buying and selling from eBay, which is one of the most successful commercial websites on the Web. Although help with eBay is provided from a single database (enabling us to control content), there are two interfaces for searching the database: My eBay Buddy and eBay Help website. Thus, similar queries will obtain the same answer from both interfaces. Figure 1 shows these two interfaces side by side.

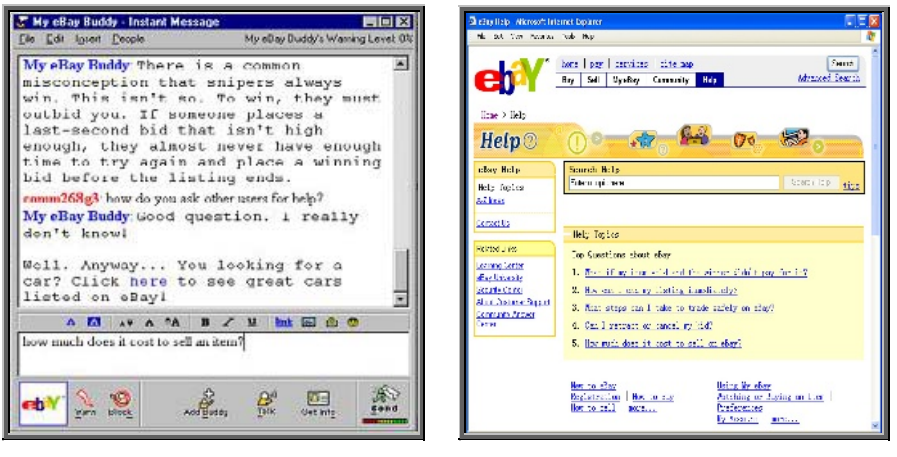

Fig. 1. Screenshots of AIM chat window with My eBay Buddy, and the eBay Help website

My eBay Buddy [17] is a natural language agent provided to AOL Instant Messenger (AIM) participants who can add My eBay Buddy to their buddy lists and chat with it to acquire information. Participants can ask open-ended questions; answers are provided by My eBay Buddy in a conversational manner. Similar agents include AgentBaseball, EllegirlBuddy, and SmarterChild.

The eBay Help website involves a classical keyword-based search paradigm, and returns results in a typical list of ranked links.

2.2 Search Tasks

Participants were given either complex or simple search tasks to find information about buying or selling items on eBay. We pre-tested the complex and simple tasks to ensure that the difficulty of each task was appropriate. The complex tasks were designed to be difficult to accomplish after a single query. In most cases participants had to search for and integrate information about two or more aspects of the eBay services. For example, one complex task was to find out how much one had to pay eBay if s/he were selling a $35 item. To come up with the right answer, participants had to learn about listing fee, final value fee, and the eBay picture service fee.

The following are four examples of the eight complex tasks:

- *Please find out if it is legal to place a bid right before the auction closes.*
- *Please find out what you should do if you don't want a certain person to bid on your item.*
- *Please find out how one can tell if a seller is reliable.*
- *Please find out what happens if the buyer doesn't pay you.*

On the other hand, the simple tasks could normally be accomplished after a single query. To ensure that the net time on task was the same for all complex and all simple tasks, there were 20 simple tasks. Pretests indicated that people completed simple tasks 2.5 times faster than complex tasks.

The following are example simple tasks:

- *Please find out what the gift icon means.*
- *Please find out what the most popular categories of eBay items are.*
- *Please find out what the PIC icon means.*
- *Please find out what the different colored stars mean.*
- *Please find out if it is possible to take back a bid.*
- *Please find out what the difference is between "proxy" and "maximum" bids.*

2.3 Participants

College students (*N*=52) from an introductory communication class participated in the experiment for course credit. All participants were native speakers of English. Participants were randomly assigned to task complexity, with gender balanced. None of the participants had ever sold anything or bought more than three items on eBay.com. Experiences with AIM were balanced across conditions. All participants signed informed consent forms upon arrival at the lab and were debriefed upon the completion of the experiment.

2.4 Procedure

The experiment was a 2x2 mixed design, with task complexity (simple vs. complex) as the between-participants factor and search interface (keyword vs. natural language search) as the within-participants factor. Participants performed all tasks in a research laboratory equipped with personal computers. Upon arrival to the laboratory, each participant was seated and assigned to a computer with both Internet Explorer and AIM.

Participants read instructions on screen. Half of them were given a list of 20 simple tasks; the other half received a list of 8 complex tasks. Half of the participants with each type of task performed the first half (10 or 4) of their tasks using My eBay Buddy (i.e., NL) and the second half (10 or 4) of tasks using eBay Help website (i.e., KW); other participants with each type of task used the two types of interfaces in reversed order.

Participants typed in an answer when they thought they had successfully accomplished a task. Upon finishing all tasks, participants responded to several questions about their search experiences. Finally, participants were asked to imagine finishing six search tasks. There were two tasks for each of three levels of complexity: high, medium and low. They were all common tasks, such as booking flight tickets, finding out the movie listing in a nearby cinema, and getting a stock quote.

2.5 Measures

Actual performance was the percentage of tasks participants successfully finished. *Perceived performance* was the percentage of tasks participants *believed* that they had successfully accomplished. *Time on task* was the average amount of time spent on each task.

Questions concerning perceived task difficulty, enjoyment of task, and confidence with search interface were asked for each task. Participants used radio buttons to indicate their responses for these questions. Each question had an independent, 10-point Likert scale. *Perceived task difficulty* was an average across task of responses to the questionnaire item, "How difficult did you feel the search was?" *Enjoyment* was an average across tasks to the questionnaire item, "How enjoyable did you find the search?" *Confidence* was an average across tasks of responses to the questionnaire item, "How confident were you with your answer?"

For each search task participants imagined to perform, participants indicated how difficult it would be when using NL and using KW. The questions were answered on 10-point Likert scales anchored by "Very Difficult" (=1) and "Very Easy" (=10).

3 Results

3.1 Manipulation Check

Time to finish the complex tasks and the simple tasks were recorded during the study. Consistent with the pre-test, each complex task took approximately three times longer to finish than did each simple task for both KW [$F(1, 24)=107.6$, $p<.001$] and NL searches [$F(1,24)=55.6$, $p<.001$; see Table 1]. There was no statistical difference for time on task between the two search interfaces ($p>.05$).

Table 1. Time on task

(minutes)	NL Mean (SD)	KW Mean (SD)
Simple	1.21(0.23)	1.33 (0.41)
Complex	3.8 (0.94)	4.17 (0.84)

3.2 Actual vs. Perceived Search Performance

Figure 2 and 3 show results for actual and perceived performance. There was a significant interaction between task complexity and search interface for both actual [$F(1,50)=26.5$, $p<.001$] and perceived [$F(1,50)=31.4$, $p<.001$] performance. Complex task participants were more successful and perceived themselves to be more successful with KW rather than NL interface. Conversely, simple task participants were more successful and also perceived themselves to be more successful with NL rather than KW interface.

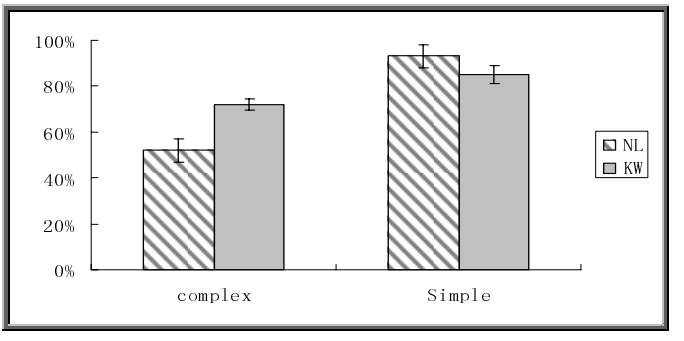

Fig. 2. Actual performance

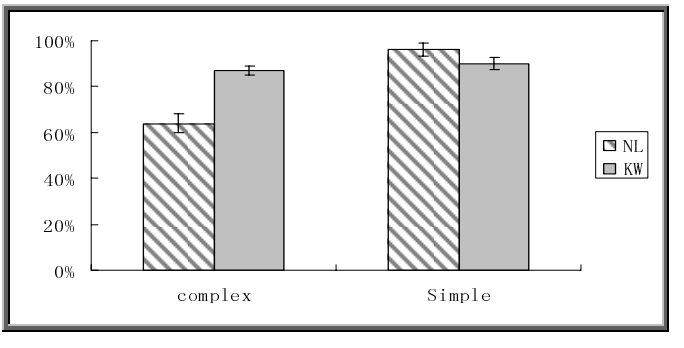

Fig. 3. Perceived performance

There was also a main effect for task complexity on performance: simple task participants had better performance than complex task participants [M_{simple}=89% vs. $M_{complex}$=62%, $F(1,50)=50.1$, $p<.01$]. Compared with complex task participants, simple task participants also believed that they had more successes [M_{simple}=93% vs. $M_{complex}$=76%, $F(1,50)=41.6$, $p<.01$].

3.3 Perceived Task Difficulty, Enjoyment of Search Experience, and Confidence with Answers

Perceived Task Difficulty. Participants' perception of task difficulty was first assessed without including performance as a covariate. There was a significant interaction effect between task complexity and search interface [F(1,50)=105.0, p<.001]. Complex task participants thought that KW made the tasks easier to perform than did NL; conversely, simple task participants thought NL was easier to work with than was KW (see Figure 4). High complexity tasks were perceived to be more difficult than were low complexity tasks (Msimple=3.20 vs. Mcomplex=4.95, F(1,50)=36.4, p<.001), but this was expected. No main effect of perceived difficulty was found for search interface [F(1,50)=.917, p>.34].

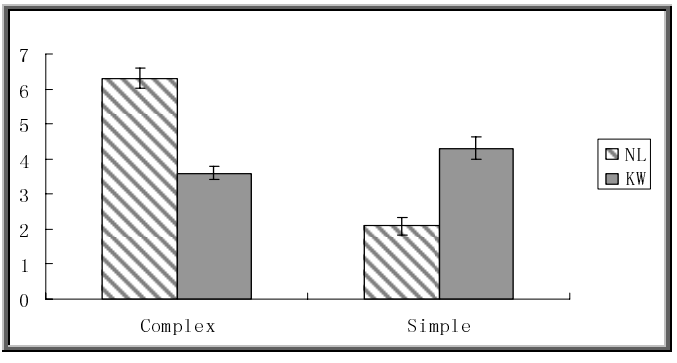

Fig. 4. Perception of difficulty

We re-examined the *perceived difficulty* data with *actual performance* as a covariate. The analysis reaffirmed our finding of an interaction effect between task complexity and search interface on perceived difficulty [$F(1,50)=28.95$, $p<.001$]. That is, even after adjusting for actual performance score, keyword search was perceived as more effective for complex tasks and less effective for simple tasks.

Enjoyment of Search Experience and Confidence with Answers. Participants' enjoyment of search experience and confidence with answers were analyzed. Table 2 presents the means and standard deviations.

There was a cross-over interaction between interface and task complexity with respect to enjoyment of search experience [$F(1,50)=47.6$, $p<.001$]. Complex task participants found using KW to be more enjoyable than using NL, while simple task participants reported the opposite. The effect remained after controlling for actual performance [$F(1,50)=19.9$, $p<.001$]. No main effect was found for search interface on participants' enjoyment of search experience [$F(1,50)=2.01$, $p>.05$]. Overall, complex task participants found the tasks to be less pleasant than simple task participants [$M_{simple}=5.87$ vs. $M_{complex}=4.75$, $F(1,50)=5.61$, $p<0.05$], and this was expected.

There was also an interaction with respect to the participants' confidence with their answers (even after controlling for actual performance). Complex task participants were more confident using KW, while simple task participants were more confident with NL interface [$F(1,50)=70.9$, $p<.001$; control: $F(1,50)=13.1$, $p<.001$].

Table 2. Perceived enjoyment and confidence

	Complex Tasks		Simple Tasks	
	NL Mean (SD)	KW Mean (SD)	NL Mean (SD)	KW Mean (SD)
Enjoyment	4.07 (1.36)	5.43 (1.59)	6.9 (1.71)	4.84 (1.83)
Confidence	5.84 (1.72)	8.03 (1.42)	8.71 (0.81)	7.65 (1.24)

Participant were more confident with their answers while working with KW search than with NL queries [$F(1,50)=8.56$, $p<.01$]. Interviews with participants during debriefing sessions concerning confidence are discussed in the following section. Not surprisingly, complex task participants felt less confident than did simple task participants [$F(1,50)=15.4$, $p<.001$].

3.4 Anticipated Search Interfaces

When participants were instructed to imagine finding answers to six search tasks with either the NL or KW interface, their indications of task difficulty confirmed our categorization of complexity, $F(2,102)=4.34$, $p<.05$. As shown in figure 5, high complexity tasks were viewed as more difficult than medium and low complexity tasks, and medium complexity tasks were viewed as more difficult than low complexity tasks [high vs. medium: $t(51)=4.06$, $p<.001$; high vs. low: $t(51)=5.17$, $p<.001$; medium vs. low: $t(51)=2.50$, $p<.05$].

There was an interaction between task complexity and choice of search interface [$F(1,50)=8.20$, $p<.01$] (Figure 5). Post-hoc tests indicated that for medium and high complexity tasks, participants preferred KW over NL, $t(51)=2.31$, $p<.05$, and $t(51)=2.62$, $p<.05$, respectively; such preference was not reported for low complexity tasks.

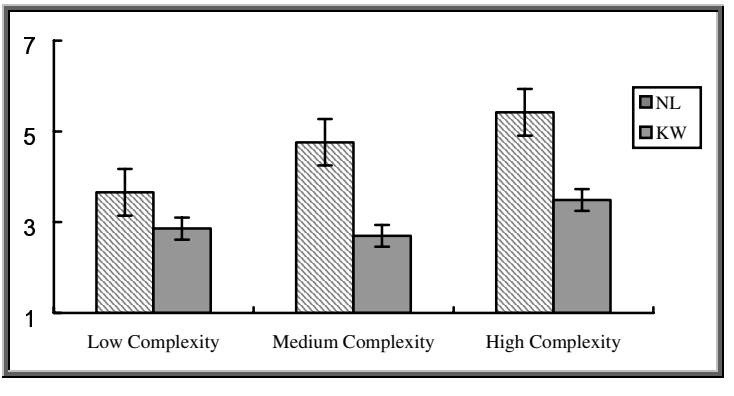

Fig. 5. Perceived difficulty for anticipated search questions

4 Discussion

The present study suggests that the selection of search interface has important and systematic effects on participant performance and perceptions. However, these effects are conditioned by the complexity of search task and the nature of queries. The general pattern we found was that NL interface was better (in both reality and perception) than KW interface for simple search tasks, while KW was better (in both reality and perception) than NL for complex search tasks. The type of search interface had no direct effect on either perceived task difficulty or enjoyment, although participants were in general more confident with KW than with NL search interface. During debriefing, several participants said that they felt more confident with KW because they had browsed pages or links that gave them peripheral affirmation. It is important to note that although there was no difference in terms of actual search performance, participants believed that they obtained more correct answers by using the KW interface than the NL interface.

4.1 Task Complexity Effect and Implications

One clear problem with NL search interfaces is that even high-quality NL systems often fail to understand what users are trying to ask. In the present study, this was true for almost all NL dialogs (i.e., with My eBay Buddy) that involved complex tasks. When My eBay Buddy could not understand the participant, it would generally ask clarifying questions. Unfortunately, the clarifying questions tend to reveal even deeper misunderstandings. Following is a sample dialogue between a participant and My eBay Buddy:

Participant: ok, tell me about your links
My eBay Buddy: What is About Me? A page that tells other eBay participants about yourself.
Participant: I'm asking for links for help
My eBay Buddy: You are asking for links for help, huh?
Participant: ahhhhhhh!
My eBay Buddy: You typed a very weird thing. What's "ahhhhhhh" supposed to mean?
Participant: I'm frustrated at you!
My eBay Buddy: For real?
Participant: for real!!!!!
My eBay Buddy: That right?
Participant: ok... where can I get more help, apart from you?

In human-human conversations, people adjust their communication strategies by asking and responding to questions in order to establish shared understandings, a process known as "grounding." [18]. When a computer agent tries to use language as a real person would, participants likely raise their expectations of the agent's ability to understand. Some previous research [19, 20] has suggested that participants may say things more freely when they have high expectations of the agent as compared to lower expectations. The freer utterances from the participant consequently impose

extra burden on natural language processing system. The gap between unconstrained participant utterances and the limited understanding capability of NL agent can lead to a spiral of failure which frustrates the participant and reduces user satisfaction.

When using the Buddy in the present study, participants with complex tasks in general found it difficult to form concise questions. As a result, they had to ask more questions than did simple task participants. Given the limited natural language processing capability of the Buddy, the more questions participants asked, the more likely the Buddy would fail to understand. Also, the more unconstrained the questions were, the more likely the Buddy would appear ignorant. It would be very useful if the agent were able to constrain user utterances by leveraging the "alignment" phenomenon observed in human-human dialogs [21].

On the other hand, simple tasks allowed the NL agent to understand and to offer relevant responses; this led to high levels of participant satisfaction in the present study. The following is a sample dialogue:

Participant: what if no confirmation email?
My eBay Buddy: Do you mean what if I submitted the registration form but didn't get a confirmation email? ……

In fact, during debriefing sessions, several simple-task participants mentioned that they thought the Buddy was smart when it asked questions back to confirm its understanding of participants' queries. Under such circumstances, it seemed to be acceptable for the agent to behave like a human being.

The above findings suggest that a natural language agent should adapt its response style to its confidence level in understanding users' input. A lower confidence level should decrease sentential responses from the system, particularly when there is no conversational history maintained by the system. Large numbers of sentential responses may mislead the participant to expect the same capability and flexibility as they would from a human interlocutor. This raised level of expectation will lead to a decreased level of satisfaction with the system when the system continues with more sentential responses even after serious misunderstandings occur.

4.2 Future Research Directions

In the present study, participants were not given the freedom to choose between the two search interfaces. One potential direction for future research is to investigate how the choice of search interface, and even the switch between search interfaces in the attempts to perform a particular task, influences user performance and perceptions. Task performance and perceptions are likely to be different when participants are able to switch from one interface to the other if they think that their initial approach is ineffective. On the other hand, this might involve duplication of effort. The research question here is whether or not combining two search interfaces/methods could provide a better user experience for information retrieval systems.

As noted earlier, it is important to understand how the response style of an NL agent influences user behaviors and attitudes. Some earlier studies have demonstrated that linguistic variations (e.g., sentence length) generated by NL agents may affect user input by soliciting alignment (i.e., mirroring) behaviors from users [19, 22].

However, linguistic alignment in human-human conversation is a bi-directional process in which two interlocutors converge. How participants would evaluate an aligning agent versus a non-aligning agent is still to be explored. On top of that, researchers must further determine when and in what ways a computer agent should align with the user to achieve or improve user satisfaction.

4.3 Final Words

The design of search methodology requires an understanding of the complex interaction between technology, psychology, and context. The present study demonstrates that any debate concerning keyword versus natural language for search must be contectualized by the complexity of search task.

References

1. Laurel, B.: Interface Agents: Metaphors with Character. In Laurel, B. (ed.), The Art of Human-Computer Interface Design. Addison-Wesley, Reading, MA (1990)
2. Doszkocs T.: Natural Language Processing in Intelligent Information Retrieval. Proceedings of the 1985 ACM Annual Conference on The Range of Computing: Mid-80's Perspective. ACM Press, 356–359
3. Guglielmo E.J., Rowe N.C.: Natural-Language Retrieval of Images Based on Descriptive Captions. ACM Transactions on Information Systems (TOIS), July 1996, Vol. 14, Issue 3, 237–267
4. Jacob, P. S., Rau, L.F.: Natural Language Techniques for Intelligent Information Retrieval. Proceedings of the Eleventh International Conference on Research & development in Information Retrieval (1988), 85–99
5. Meng, F.: A Natural Language Interface for Information Retrieval from Forms on the World Wide Web. Proceeding of the 20th International Conference on Information Systems (1999), 540–545
6. Turtle H.: Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (1994), 212–220
7. Chen, H., and Dumais, S.: Optimizing Search by Showing Results in Context. Proceedings of CHI 2001. ACM Press, 145 – 152
8. Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R.: Using Latent Semantic Analysis to Improve Access to Textual Information. Proceedings of CHI '88. ACM Press, 281 – 285
9. Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J., & Pirolli, P.: Using Thumbnails to Search the Web. Proceedings of CHI 2001. ACM Press, 198 – 205
10. Borgman, C. L.: The Participant's Mental Model of an Information Retrieval System: An Experiment on a Prototype Online Catalog. International Journal of Man-Machine Studies, 24(1) (1986), 47-64
11. Borgman, C. L.: Psychological Research in Human-Computer Interaction. In M. Williams (Ed.), Annual Review of Information Science and Technology, Vol. 19. White Plains, NY: Knowledge Industry Publications (1984), 33-64
12. Greene, S.L., Devlin, S.J., Cannata, P.E. and Gomez, L.M.: No IFs, ANDs or ORs: a Study of Database Querying. International Journal of Man-Machine Studies, 32, 3 (1990), 303-326

13. Cleverdon, C.W.: The Significance of the Cranfield Tests on Index Languages. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan (Eds). Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, (1991), 3 – 12
14. Cooper, W. S.: Getting Beyond Boole. Information Processing & Management 24(3) (1988), 243-248
15. Muramatsu, J., Pratt, W.: Transparent Queries: Investigating Participants' Mental Models of Search Engines. SIGIR: Proceedings of the 24th International Conference on Research & Development in Information Retrieval (2001)
16. Navarro-Prieto, R, Scaife, M, and Rogers, Y.: Cognitive Strategies in Web Searching. Proceedings of the 5th Conference on Human Factors & the Web (1999)
17. Conversagent, Inc. http://www.conversagent.com
18. Clark, H.H.: Using Language. Cambridge University Press, New York (1996)
19. Brennan, S.: Conversation with and through Computers. Participant Modeling and Participant-Adapted Interaction, 1 (1991), 67–86
20. Pearson, J., Pickering, M.J. Branigan, H.P., McLean, J.F., Nass, C.I., Hu, J.: The Influence of Beliefs about an Interlocutor on Lexical and Syntactic Alignment: Evidence from Human-Computer Dialogues. Proceeding of the 10th Annual Architectures and Mechanisms of Language Processing Conference (2004)
21. Nass, C.I., Hu, J., Pearson, J., Pickering, M.J., Branigan, H.P.: Linguistic Alignment in HCI vs. CMC: Do Participants Mimic Conversation Partners' Grammar and Word Choices? Unpublished manuscript (2004).
22. Zoltan-Ford, E.: How to Get People to Say and Type What Computers Can Understand. International Journal of Man-Machine Studies, 34(4) (1991), 527-547