Rutgers' TREC 2001 Interactive Track Experience

N.J. Belkin, C. Cool*, J. Jeng, A. Keller, D. Kelly, J. Kim, H.-J. Lee, M.-C. Tang, X.-J. Yuan
School of Communication, Information & Library Studies
Rutgers University
New Brunswick, NJ 08901-1071
*GSLIS, Queens College, CUNY
nick@belkin.Rutgers.edu ccool@qc.edu [judyjeng | amkeller | diane | jaykim | hyukjinl | muhchyun | xjyuan]@scils.rutgers.edu

Abstract

Our focus this year was to investigate methods for increasing query length in interactive information searching in the Web context, and to see if these methods led to changes in task performance and/or interaction. Thirty-four subjects each searched on four of the Interactive Track topics, in one of two conditions: a "box" query input mode; and a "line" query input mode. One-half of the subjects were instructed to enter their queries as complete sentences or questions; the other half as lists of words or phrases. Results are that: queries entered as questions or statements were significantly longer than those entered as words or phrases (twice as long); that there was no difference in query length between the box and line modes (except for medical topics, where keyword mode led to significantly more unique terms per search); and, that longer queries led to better performance. Other results of note are that satisfaction with the search was negatively correlated with length of time searching and other measures of interaction effort, and that the "buying" topics were significantly more difficult than the other three types.

1 Introduction

The goal of the TREC 2001 Interactive Track was that the participants in the Track carry out exploratory studies which could lead to testable hypotheses (or firm research questions) to be investigated in the course of the TREC 2002 Interactive Track. These exploratory studies were to be carried out by having subjects search on a variety of predetermined topics on the "live" Web. At Rutgers, we decided to focus primarily on the issue of query length in interactive searching, with secondary interests in subject use of a feedback device, and in the effect of highlighting of query terms in search results.

We were interested in query length for three reasons. The first of these is the well-known finding that, for best-match retrieval engines, the longer the query, the better the retrieval results. Since it is also well-known that users of Web search engines typically enter rather short queries, we were interested in methods that might increase query length. The second reason is that our work was also connected with the NSF-funded MONGREL project in which we are collaborating with colleagues at the University of Massachusetts, Amherst (MONGREL). This project is concerned with using language-modeling methods (e.g. Ponte & Croft, 1998) for developing topic and user models; for this purpose, it is important to have fairly long queries. The third reason for considering query length was that the Interactive Track topics were designed to be of four different "types": medical; travel; buying; and project, and that the topics associated with these types were often couched (or could be couched) as questions. We hypothesized that differences between these types might show up in either length of query for each, or in framing of question for each. In either case, longer queries should enhance the chances of discovering any such differences.

We considered two different methods for increasing query length. The first was to vary the size and format of the query input mode. Karlgren & Franzén (1997) found that subjects who were asked to input queries in a box-like query input window (one in which the input query was wrapped for multiple lines) had significantly longer queries than subjects who entered queries in a standard Web-browser query input line. We decided to test this result in our study, which had more subjects than they did, and which also had a greater variety of search topic types. Our hypothesis was that the box mode would lead to longer queries than the line mode, for two reasons. The first is that the perceived space for query entry is larger in the box mode; the second is that the entire query, no matter how long (within some reasonable limits) would be visible in the box mode, and therefore people would be encouraged to continue query entry. The second method of increasing query length was to vary the form of query. We did this by instructing subjects either to enter their queries as complete questions or sentences, or to enter their queries as a list of words or phrases. Our hypothesis was that the former would lead to longer queries than the latter.

We were also secondarily interested in studying use of feedback facilities in Web searching, following up on our previous TREC Interactive Track studies (cf. Belkin, Cool, et al. 2001). This was implemented in our system this time as a "copy-and-paste" facility for moving text from displayed pages directly into the query. Finally, we decided to consider the perceived usefulness of highlighting of query terms.

2 System

Searching was conducted through a proxy server and our own interface, using the Netscape browser, to the Excite search engine. Our initial interface consisted of a query input window, which was either a standard 50-character line, or a scrollable 40-character by five line box, in which input text was automatically wrapped, and a "search" button. The query was displayed at the top and bottom of each retrieved Web page (result or linked), along with a query modification window, into which text from the page could be copied, and then copied into the query and run as a modified query. All query terms were highlighted in query result lists and in viewed pages. All displayed, visited and printed pages were logged, as were all queries and query modifications. Screen shots of the interface are available at http://www.scils.rutgers.edu/mongrel/trec.html.

3 Methods

3.1 Design

The Interactive Track specification provided sixteen search topics, four topics for each of four different topic "types": medical; travel; buying; project. Within each type, there were two "fully-specified" topics, and two "partially-specified" topics. Our study was designed with one within-subjects factor (line vs. box query input mode), and one between-subjects factor (complete question/sentence vs. list of words/phrases). In order to obtain adequate representation on all topic types and on the specified-partially specified dimension, we needed to have 32 subjects (in fact, we ran 34 subjects, duplicating the first two subject conditions), sixteen in the group instructed to search using a complete question/sentence; sixteen in the group instructed to search using a list of words or phrases. Each of the subjects searched on four topics, the first two fully-specified, the second two partially-specified. This order was determined on the basis that it would be easier for the subjects to do the fully-specified topics. Each subject performed one specified and one partially-specified search using the box input mode, and one specified and one partially-specified search using the line input mode. Search time was limited to a maximum of fifteen minutes. The query input modes were alternated, and the order in which they were performed was systematically varied for the entire group of subjects. The design of the study is shown in Table 1, where **Snn** is the subject number, column one defines the combination of type of query and order of input mode, and each cell represents the topics and the order in which they were searched by each subject. The order in which subjects were run is indicated by highlighting in Table 1, with the diagonal pattern continued, first recurrence beginning with S03.

Condition **Q** Order **LB**	**S01** Medical 1 Buying 3 Project 15 Travel 14	**S02** Medical 2 Travel 6 Project 16 Buying 11	**S03** Buying 3 Travel 5 Medical 9 Project 15	**S04** Buying 4 Project 8 Travel 14 Medical 10	**S05** Travel 5 Project 7 Buying 12 Medical 9	**S06** Travel 6 Medical 2 Buying 11 Project 16	**S07** Project 7 Medical 1 Travel 13 Buying 12	**S08** Project 8 Buying 4 Medical 10 Travel 13
Condition **Q** Order **BL**	**S09** Medical 1 Buying 3 Project 15 Travel 14	**S10** Medical 2 Travel 6 Project 16 Buying 11	**S11** Buying 3 Travel 5 Medical 9 Project 15	**S12** Buying 4 Project 8 Travel 14 Medical 10	**S13** Travel 5 Project 7 Buying 12 Medical 9	**S14** Travel 6 Medical 2 Buying 11 Project 16	**S15** Project 7 Medical 1 Travel 13 Buying 12	**S16** Project 8 Buying 4 Medical 10 Travel 13
Condition **T** Order **LB**	**S17** Medical 1 Buying 3 Project 15 Travel 14	**S18** Medical 2 Travel 6 Project 16 Buying 11	**S19** Buying 3 Travel 5 Medical 9 Project 15	**S20** Buying 4 Project 8 Travel 14 Medical 10	**S21** Travel 5 Project 7 Buying 12 Medical 9	**S22** Travel 6 Medical 2 Buying 11 Project 16	**S23** Project 7 Medical 1 Travel 13 Buying 12	**S24** Project 8 Buying 4 Medical 10 Travel 13
Condition **T** Order **BL**	**S25** Medical 1 Buying 3 Project 15 Travel 14	**S26** Medical 2 Travel 6 Project 16 Buying 11	**S27** Buying 3 Travel 5 Medical 9 Project 15	**S28** Buying 4 Project 8 Travel 14 Medical 10	**S29** Travel 5 Project 7 Buying 12 Medical 9	**S30** Travel 6 Medical 2 Buying 11 Project 16	**S31** Project 7 Medical 1 Travel 13 Buying 12	**S32** Project 8 Buying 4 Medical 10 Travel 13

Table 1. Subject assignment form. Q = question/sentence T = word/phrase L = line input B = box input. Specified topics are numbers 1-8; partially specified topics are numbers 9-16.

3.2 Procedure

Volunteer subjects were recruited primarily from the population of students at the School of Communication, Information and Library Studies (SCILS) at Rutgers University. The recruitment notice specified that the single session for which they were volunteering would last about two hours. The search sessions were held at the Information Interaction Laboratory at SCILS, which allows unobtrusive video and audio recording of searching behavior. Upon arrival, subjects completed first an Informed Consent form, and then a brief demographic questionnaire eliciting age, gender, educational background, and a variety of measures of previous searching experience and searching attitudes. They were then given a general description of the tasks that they would be asked to perform during the experimental session. Then they were handed a specification of the first search topic, on a form which asked them to indicate whether they knew the answer to the search topic, or where to find an answer, and their confidence in that judgment. Then they went to the search station, and began their search on the first topic. Subjects were instructed to "think aloud" as they searched, and their thinking aloud, as well as the monitor while searching were recorded on videotape. Subjects were instructed to print out all pages which helped them to answer the search topic. They were told that they could search for up to fifteen minutes, but could quit searching as soon as they felt they were done. On completion of the search, they answered a brief questionnaire about that search experience, and then explained to the experimenter present why they printed out each page that they did (i.e., what it was about that page that helped them to answer the search question/topic). This procedure was continued for all four search topics. After the fourth topic cycle, subjects were administered an exit interview, which was recorded on audio tape, eliciting their opinions about the different query input modes, about the query type that they were asked to use, about the query modification and highlighting features, and about the general characteristics of the systems that they used, as compared to those they ordinarily use. Examples of the data collection instruments are available at http://www.scils.rutgers.edu/mongrel/trec.html

3.3 Subjects

The subjects for this study were primarily students in the Masters of Library and Information Science program at SCILS, but also included some undergraduate students in communication courses. Of the 34 subjects, 5 were male, 29 female. The age distribution 44% between 20 and 29 years, 30% between 30 and 39 years, 12% between 40 and 49 years, and 14% over 50 years.

4 Results

4.1 Query and interaction characteristics

Queries were characterized according to the following measures: number of queries per search; average query length (in words) per search; number of unique query terms per search. Interaction was characterized according to the number of unique pages seen (i.e. urls displayed) and the number of unique pages viewed (i.e. opened by following a link). The data for these measures, for all searches, are displayed in Table 2.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
unique seen	133	10	83	23.28	16.21
unique viewed	133	0	16	3.65	2.61
number of queries	134	1	8	2.13	1.70
number of unique terms in search	133	1	28	7.23	4.72
AVLENGTH	134	1	17	5.54	3.29
Valid N (listwise)	132				

Table 2. Query and interaction measures for all searches

When the data were analyzed to see if the query type (i.e. question/sentence vs. list of words/phrases) affected query characteristics or interaction characteristics, we found that the two measures of query length, number of unique terms in the search, and average query length, were significantly greater for the question/sentence type, using the t test (unique terms in search, $t(131) = 9.14$, $p<.01$; average query length, $t(132) = 11.94$, $p<.01$). The data for all searches on all topics are displayed in Table 3. This relationship held for all different topic types, when analyzed separately, and also for both specified and unspecified queries (with the exception of the medical topics, see below).

Group Statistics

	condition	N	Mean	Std. Deviation	Std. Error Mean
unique seen	0	71	23.31	16.04	1.90
	1	62	23.24	16.53	2.10
unique viewed	0	71	3.70	2.28	.27
	1	62	3.58	2.97	.38
number of queries	0	71	2.30	1.84	.22
	1	63	1.94	1.51	.19
number of unique terms in search	0	71	9.97	4.48	.53
	1	62	4.10	2.53	.32
AVLENGTH	0	71	7.76	2.90	.34
	1	63	3.02	1.31	.17

Table 3. Query and interaction characteristics for question/sentence (0) and list of words/phrases (1) query types.

There was no significant difference in query length, or any other query or interaction characteristic between the box and line query input mode when all searches for all topics are considered. However, when each type of topic was considered separately, an interesting difference became apparent. For the medical topic type, the number of unique terms per search, and the number of queries per search, were significantly greater in the line mode than in the box mode (unique terms in search, $t(31) = 2.41$, $p<.05$; number of queries, $t(31) = 2.82$, $p<.01$). The data for the various measures for box vs. line mode are displayed in Table 4. It is interesting to note that the number of queries in the line condition, and the number of unique terms in the line condition, are both about double those in the box condition. Although the average query length for the line condition is somewhat longer than for the box condition, this difference is not statistically significant. But when considering specified vs. partially-specified medical topics, average query length is significantly longer (mean specified = 6.34; mean partially-specified = 4.32, $t(31)=2.06$, $p<.05$).

Group Statistics

	size of search box	N	Mean	Std. Deviation	Std. Error Mean
number of queries	Line	16	3.19	2.56	.64
	Box	17	1.53	1.18	.29
number of unique terms in search	Line	16	10.19	6.67	1.67
	Box	17	5.29	2.57	.62
AVLENGTH	Line	16	6.28	3.19	.80
	Box	17	4.50	2.49	.60
unique seen	Line	16	27.63	18.65	4.66
	Box	16	17.19	9.70	2.42
unique viewed	Line	16	4.31	2.70	.68
	Box	16	2.56	2.19	.55

Table 4. Query and interaction characteristics for medical topics in line and box modes.

4.2 Performance of task

Performance, that is, correct and complete answering of the topic, was measured by the experimenters on the basis of the pages printed out by each subject. We did this by deciding whether the pages which were printed out either: did not respond to the topic/question at all; partially answered the topic/question; or completely answered the topic/question. Although we also had a measure of performance based on the subjects' assessment after each search, we used our measure in preference because we sometimes found what appeared to be misunderstandings of the tasks

on the part of the subjects. Using this measure, we found some interesting results. Figure 1 displays the average length of the query against performance on the task, for all searches. Here we see that as the query gets longer, performance regularly becomes better, although this is a descriptive finding, only.

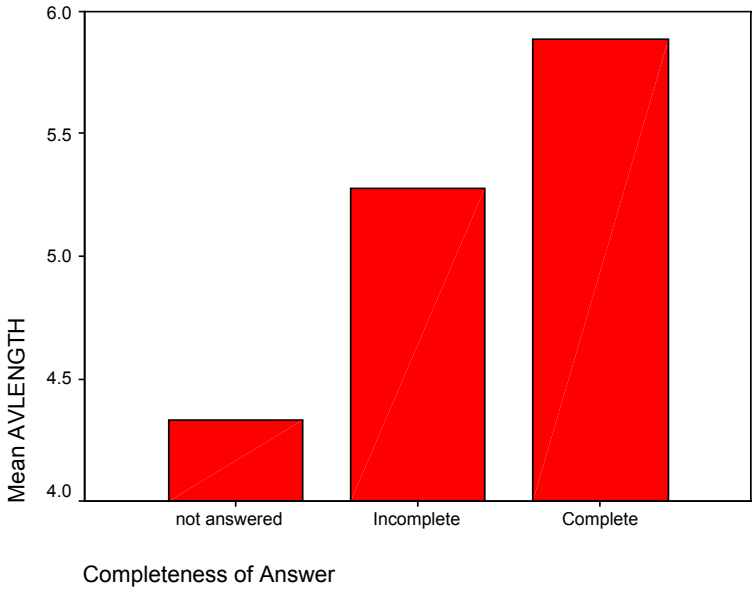

Figure 1. Task performance versus average query length.

Performance was not significantly related to either input mode or query type, nor, in general, to any other query or interaction characteristics. However, when analyzing performance by topic type, the buying topics turned out to be significantly more difficult than the others. These data are displayed in Table 5; the chi-squared test gives χ^2 (6) = 14.89, $p<.05$.

TASKTYPE * completeness of answer/objective Crosstabulation

Count

		completeness of answer/objective			
		not answered	partially answered	completely answered	Total
TASKTYPE	medical		15	16	31
	travel		13	17	30
	Project		12	19	31
	buying	3	20	9	32
Total		3	60	61	124

Table 5. Performance versus topic type.

4.3 Satisfaction with search

After each search, subjects were asked to indicate their degree of satisfaction with that search on a five-point Likert scale, with 1 indicating unsatisfied, and 5 completely satisfied. Using similar scales, they were also asked to indicate their familiarity with the task (i.e. topic) and whether previous knowledge of the topic had helped them in their search. These measures were correlated with the query and interaction measures indicated before, and with one another. The most interesting results from this analysis are that the major interaction measures, number of queries, unique seen and unique viewed are moderately, but significantly negatively correlated with satisfaction (using the Pearson correlation, these correlations are: number of queries -.350**; unique seen -.314**; unique viewed -.182*;

where **= significance at .01, *=significance at .05). A further interesting relationship is that previous knowledge of the topic is significantly correlated with satisfaction (.380*)

4.4 Query modification and highlighting

There was very little use of the query modification feature by our subjects. Overall, only 9 subjects (26%) used this feature at all. Five of them used it two or more searches, and the feature was used in only 16 (12%) of all of the searches. We asked about the usefulness of the query modification feature in the exit interview, and about the subjects' familiarity with this type of feature. The narrative data are still being analyzed, but we have some preliminary results. Almost all subjects were unfamiliar with this feature (rating of 1 or 2 on a 5-point Likert scale where 1 is completely unfamiliar and 5 is completely familiar, M = 1.72), and almost all rated it as not useful (1 or 2 on a 5-point Likert scale where 1 is useless and 5 very useful, M = 2.03). In general, those who who gave the feature low usefulness scores were also unfamiliar with the feature. The few subjects who gave a high usefulness rating also claimed to have high (4,5) familiarity with this type of feature. When asked to explain the reasons for their usefulness ratings, those who gave negative ratings had three types of reasons, as follows. (1) They didn't see a need for the feature, since it was just as easy to type directly; (2) they were used to doing things differently; and (3) they were unfamiliar with the feature.

We investigated the usefulness of highlighting also by asking about it in the exit interview. As with the query modification feature, the narrative data are still being analyzed, but we present preliminary results. The mean usefulness rating was 4.21, on the same 5-point scale as for usefulness of the query modification feature. This high rating was consistent across all topic categories and subject conditions. Reasons given for high ratings (4,5) were that highlighting made it easier to and quicker to tell when something was relevant, and also to tell when something was not relevant. Reasons give for low ratings (1,2) were that a lot of irrelevant words were highlighted, and that there was confusion between highlighted query terms and links to other pages.

5 Discussion

On the basis of our results, it appears that encouraging people to enter their queries as questions or complete statements will lead to longer queries, which in turn will lead to greater satisfaction with the search, and to better search performance. The strongest result that we obtained with respect to query length was that question/sentence queries were significantly longer than keyword queries. This is certainly at least in part due to the inclusion in the former of words and phrases which are traditionally found on stop lists, and are discouraged or not used in keyword queries (cf. Toms et al., this volume). However, given that the general task that was required of all of the topics was very close to a complex question-answering task, we believe that the inclusion of such words in a query will lead to better performance in systems which are explicitly designed to support this type of task. For instance, all of the systems in the TREC Question and Answer Track make some use of such words (e.g. when, where, what, how) in order to classify the type of question and thereby increase performance. Although the system used in our study, Excite, does not directly support this task, we find it noteworthy that performance did nevertheless increase with query length.

We were unable to replicate Karlgren and Franzen's (1997) results with respect to obtaining longer queries using the box input mode than the line input mode. Our initial hypothesis regarding query input mode, that the box input mode would lead to longer queries than the line input mode, was not directly supported by our results. This failure deserves further analysis of our qualitative data, where subjects were asked to explain their attitudes toward, and preferences for the different query types and input modes. However, we did notice that overall our queries, for both input modes, were slightly longer than what has been reported in the web searching literature. Jansen, Spink and Saracevic (2000) analyzed 51,473 Excite queries and found that on average, the queries contained 2.21 terms. Silverstein, Henzinger, Marais and Moricz (1999) studied over one billion queries from Altavista and found that 72.4% of the queries contained 2 or fewer terms. But our line mode queries were 3.02 words long. In looking for an explanation for these longer queries, we noticed that our line input mode was 50 characters long. This was 32 characters longer than the line input mode in Karlren and Franzen's (1997) study! We conducted a follow-up survey[1] of seven web search services' query input boxes, including Excite and Altavista, in order to better understand the relationship between the size of query input facilities and query length. The size of each search services' query box, measured in characters, was as follows: Altavista, 30; Excite, 37; Google, 35; InfoSeek, 30; NorthernLight, 30; WebCrawler, 33; Yahoo, 30. In only three of these query boxes (Altavista, Google and InfoSeek) could one enter a query that exceeded the number of displayed characters. Thus, we conclude that both our line input mode, and our box input mode encouraged longer queries. This may, perhaps, provide some support for our initial

[1] Survey conducted on 28 January 2002.

ideas about query length and query input size, and at least in part explain why our results did not replicate those of Karlgren and Franzén (1997).

The results for the medical topics with respect to numbers of queries and unique query terms per search in line versus box mode are quite strange, and we have not yet come up with a convincing explanation. Since the average length of each query is about the same in the two conditions, it appears that the increase in the unique number of search terms is due to the increase in the number of queries in the line condition. But we have no reasonable explanation of why this should increase for the line mode only in the medical topics. We considered the possible influence of long words being associated with medical topics, but found no difference in average word length between medical topics and the other topic types.

Performance needs to be better studied. The positive relationship between average query length and performance was encouraging, but how this happens, and in particular how it relates to measures of satisfaction and measures of interaction needs to be clarified. Also, it appears that the actual tasks that the subjects engage in need to be better defined, so that misunderstanding is less likely to happen. Examples of obvious misunderstanding of the task led us to determine performance not according to the subjects' own evaluation with respect to the task, but rather by our own understanding of the task, and of their answers. This is clearly not ideal, and needs to be addressed in future studies of this sort.

The negative correlations between interaction measures and satisfaction could be explained either because the subjects had to do a lot of interaction to get the answer, or because they did a lot of interaction, and still didn't get the answer. The relationship between familiarity (prior knowledge) of the topic and satisfaction is of some interest as well, since neither of these factors appears to have had an effect on performance. This, of course, may be due to our particular performance measure.

The lack of use of the query modification feature has several possible explanations. It was somewhat cumbersome to use, it was not very easy to find, and the idea was not very familiar to the subjects. It is also the case that it might not have been well-suited to the task at hand, especially given its time constraints. Despite these caveats, we think that these results confirm what others, ourselves included, have found; that explicit feedback mechanisms are in general too peripheral to searching tasks to be taken up in a major way by end users.

6 Conclusions

A goal of the TREC 2001 Interactive Track was that the studies conducted in it should lead to hypotheses which could be tested (or research questions which could be investigated) in the TREC 2002 Interactive Track. Our results in this year's Interactive Track raise a good number of issues which should be further investigated, in particular those having to do, on the one hand, with the negative relationship of our measures of degree of interaction with subject satisfaction, and on the other hand, with the relationship of query length to task performance. To address these two general problem areas, we will need to: develop new interaction measures and/or design our study to better measure interaction; develop techniques which we believe will lead to reduced interaction; investigate new methods for increasing query length; study the effect of "non-content-bearing" words in queries on task performance; and, develop some better measure(s) of performance, or at least design the new studies to better allow measurement of task performance.

In our TREC-9 study (Belkin, Keller et al., 2001), we developed and tested an interface which displayed, in two rows of six scrollable panes, the texts of the retrieved documents, beginning at the "best passage" in that document (our MDD interface). We believe that a large part of the interaction that we observed in TREC 2001 searches had to do going back and forth between search result lists and the actual pages to which the results pointed, and that searching through those pages also increased interaction effort. This leads us to our

> **Hypothesis 1**: Displaying search results as in our TREC-9 MDD system will increase user satisfaction over displaying search results as lists of references or links to the full texts of the retrieved documents.

Although we observed a consistent trend of increase in task performance with increased query length, this result is somewhat clouded by the measure of performance that was used, and by uncertainty regarding what aspects of query length led to this result. Thus, we will study methods of increasing query length, and design our TREC 2002 Interactive Track study to investigate the effects of different query types and features on performance, and will test our

> **Hypothesis 2**: Increased query length leads to better task performance.

7 Acknowledgements

This research was funded in part by Grant Number IIS 99-11942 from the National Science Foundation. We again gratefully thank our volunteer subjects, without whose collaboration none of our work could be done.

8 References

Belkin, N.J., Cool, C., Kelly, D., Lin, S.-J., Park, S.Y., Perez-Carballo, J., Sikora, C. (2001) Iterative exploration., design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing and Management, 37*, 403-434.

Belkin, N.J., Keller, A., Kelly, D., Perez-Carballo, J., Sikora, C., Sun, Y. (2001) Support for Question-Answering in Interactive Information Retrieval: Rutgers' TREC-9 Interactive Track Experience. In: E.M. Voorhees & D. K. Harman (eds.) The Ninth Text REtrieval Conference (TREC-9) (pp. 463-475). Washington, DC: GPO.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management, 36*(2), 207-277.

Karlgren, J. & Franzén, K. (1997) Verbosity and interface design. http://www.ling.su.se/staff/franzen/irinterface.html

MONGREL. Web page for the MONGREL Research Project. http://scils.rutgers.edu/mongrel

Ponte, J. M. & Croft, W.B. (1998) A language modeling approach to information retrieval. In: *SIGIR '98*. New York: ACM, 275-281.

Silverstein, C., Henzinger, M., Marais, J., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6-12.