University of North Carolina's HARD Track Experiments at TREC 2004

Diane Kelly, Vijay Deepak Dollu, & Xin Fu

School of Information and Library Science
University of North Carolina at Chapel Hill
100 Manning Hall, CB #3360
Chapel Hill, NC 27599-3360

[dianek | vijayd | fu] @ email.unc.edu

Abstract

In the experiment described in this paper, we investigate the effectiveness of a document-independent technique for eliciting additional information from searchers about their information problems. We propose that such a technique can be used to elicit terms for use in query expansion and as a follow-up when ambiguous queries are initially posed by searchers. We use a clarification form to obtain additional information from searchers and create a series of experimental runs based on the information that we obtained from this form. Although we were successful at eliciting more information from searchers, we were unable to demonstrate that this additional information increased performance because of an indexing error that resulted in very poor performance for our baseline and experimental runs. Additionally, we use our clarification form to investigate an alternative measure of topic familiarity and demonstrate how it relates to the length of searchers' topic descriptions and responses to our clarification form.

1 Introduction

In this year's HARD track, we took advantage of the one-shot interaction provided by the clarification form to investigate the effectiveness of various techniques for eliciting additional information from searchers about their information problems. Our experiments were motivated by several interests. First, we were interested in creating a feedback technique for use in situations where a searcher's initial query is unclear or ambiguous. Although previous research has successfully developed and evaluated the clarity measure for predicting query ambiguity [3], it remains unclear what steps should be taken to clarify ambiguous queries once they are identified. Thus, we sought to extend the work on the clarity measure by investigating techniques that could potentially be used to follow-up ambiguous queries.

Second, we were interested in developing and evaluating a generic, document-independent feedback technique that could be used in multiple information-seeking situations. This interest was motivated by the supposition that traditional relevance feedback techniques, which typically present top-ranked documents or keywords to searchers for feedback, are unlikely to work well in situations where ambiguous queries are posed because there is a large chance that documents retrieved in response to such queries will be irrelevant. Further, a generic, document-independent feedback form could potentially be used to assist digital reference librarians elicit more information from patrons about their information problems and manage voluminous service requests by providing support for question triage. For instance, when initial requests are received by a digital reference service the clarity of the request could be determined. If the request is determined to be inadequate, then the patron could be asked to complete a generic clarification form with the goal of eliciting more information from the patron. Although we are proposing a generic clarification form, it is possible to imagine digital reference clarification forms tailored to specific topic areas and collection types.

Finally, we took advantage of the experimental setup of the HARD track to investigate our interest in developing techniques for measuring searchers' topic familiarity. The familiarity data collected as part of the topic metadata provided a nice opportunity in which to compare data collected using an alternative familiarity measure. Thus, we included a familiarity measure as part of our clarification form. One motivation for including this question as part of our clarification form was that during the planning stages of this year's track, there was quite a lot of discussion about appropriate techniques for measuring familiarity. Although some consensus was reached about how this would occur in this year's track, there is still much work to be done on developing valid and reliable techniques for assessing familiarity. Given the length and complexity of the discussion surrounding familiarity in the planning stages of this year's track, we hoped that our results would provide some insight into this complex issue, and perhaps prove useful in the track's future planning.

2 HARD 2003 Clarification Forms

In last year's HARD track [1], the clarification form was the major technique used to collect data for use in experimental runs. In designing our clarification form we first considered techniques used by sites in last year's HARD track, the results of these techniques, and the feedback provided by last year's searchers. Among the 12 HARD sites whose papers were collected in the 2003 TREC Proceedings, 10 used clarification forms as one of their experimental techniques. The essential purpose of the clarification form was to elicit additional information, usually through relevance feedback, from users about their topics. The information elicited from the clarification forms was used for a variety of experimental techniques, although most often, query expansion and document re-ranking. In total, the 10 sites submitted 21 forms for each topic; most contained more than one question.

There were two general approaches to generating clarification forms in last year's track. The first of these was to use some sort of document surrogate, which had been retrieved in response to searchers' baseline queries, to populate the clarification form. Searchers were shown these surrogates, which typically consisted of terms/phrases, sentences and passages (including headlines) and asked to mark them in some way. The number of surrogates displayed on each form varied from team to team, as did the methods used to extract the surrogates. Six teams employed clarification forms that displayed terms and/or phrases to searchers, three teams used sentence surrogates in their clarification forms and three teams presented document passages to searchers. In several cases, the clarification form presented two or more types of surrogates to searchers for evaluation. Searchers were most often provided with one of two form elements to provide feedback: check boxes or radio buttons. For example, check boxes were often provided next to surrogates and searchers were instructed to check all relevant items. Alternatively, a series of radio buttons corresponding to each surrogate or surrogate-cluster and displaying labels with different relevance values (e.g. relevant, not relevant, not sure), were provided and searchers were asked to select one value for each item or group of items.

The second approach to generating clarification forms was to present searchers with questionnaire-type items whose content was not generated from initial search results. These items included those related to searchers' previous searching experiences and general preferences, as well as those that asked searchers to enter additional key terms describing their topics. For instance, one team probed searchers' recent searching experience, preference for sub-collections and time frame for documents. Quite often, clarification was sought on specific aspects of the metadata. For instance, one team asked searchers to choose their preferred information level (overview versus details) and nature of results (documents versus answers). Numerous sites used an open-ended question to elicit additional relevant terms from searchers; about half of the clarification forms asked searchers to provide additional terms by entering them into a scrollable textbox. Most sites who used this technique reported positive results.

Finally, we considered the feedback that HARD 2003 searchers provided at the LDC HARD website (http://www.ldc.upenn.edu/Projects/HARD/cfs.html) as an important resource to inform the design of our clarification form. In particular, we observed that last year's searchers preferred to have a text box in which to include additional relevant terms over other methods, and they enjoyed having a large enough space in which to specify more information about their topic. Thus, as much as possible, we tried to incorporate this feedback into the design of our clarification form.

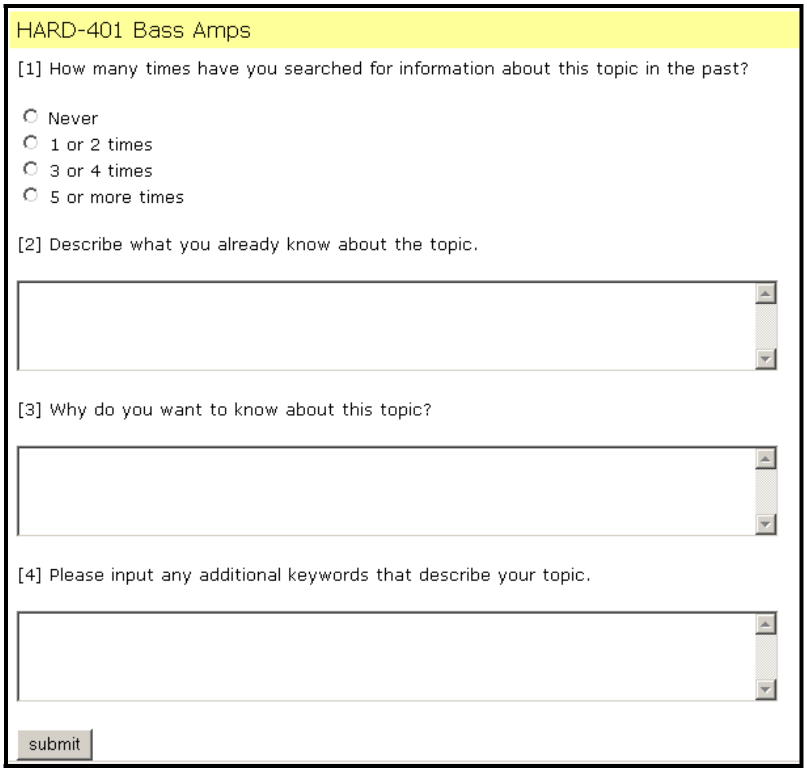
3 HARD 2004 Clarification Form

Based on our review of previous approaches and our particular research interests, we designed a clarification form which consisted of four questions, and that could be used for all topics without modification. This clarification form is displayed in Figure 1. The first question that we presented was a familiarity question, which asked searchers to indicate how many times they had searched for information about their topics in the past. Searchers were provided with four choices: (1) never; (2) 1 or 2 times; (3) 3 or 4 times; and (4) 5 or more times. We were motivated to use this particular question to assess familiarity because we were interested in understanding how familiarity might be inferred from a searcher's behavior.

If one were able to track a searcher's behavior over time, then this type of measure might prove useful, given that it is, in fact, related to a searcher's actual topic familiarity.

Questions 2, 3 and 4 were designed to elicit information from searchers about their topics. In designing clarification form features to elicit this information, we were careful to use large text boxes that allowed users to view the entirety of their responses and hopefully, as found in previous studies [2, 5], encourage them to type in longer responses than they would if presented with a short line. Questions 2 and 3 were open-ended questions (although 2 is presented as a statement), and encouraged searchers to respond in natural language. Question 2 asked searchers to describe what they already know about the topic, and Question 3 asked searchers to indicate why they want to know about the topic. Our goal in using these questions was to encourage searchers to talk more about their topics, and hopefully in doing so, have them provide additional information that might prove useful in retrieval. Our selection of these two questions was based on an examination of previous research on face-to-face reference interviews [c.f. 4] and reference textbooks which describe best practice [c.f. 6].

Question 4 asked searchers to list any additional keywords that describe their topics. As mentioned earlier, a number of participating groups from last year's track used a question like this, some with quite successful results. Thus, we included this on our form with hopes that it would again provide some useful data. It was also the case that the majority of clarification forms from last year asked searchers to make a selection of good terms from a list of extracted terms. Thus, we further hoped that this question would allow us to take advantage of the priming that searchers might receive by being exposed to such lists of terms before they completed our form in the experimental rotation. The assumption, of course, is that if searchers see a good term on another clarification form, then there is a possibility that they will remember and enter it when they reach our form.



HARD-401 Bass Amps

[1] How many times have you searched for information about this topic in the past?

○ Never
○ 1 or 2 times
○ 3 or 4 times
○ 5 or more times

[2] Describe what you already know about the topic.

[3] Why do you want to know about this topic?

[4] Please input any additional keywords that describe your topic.

Figure 1. UNC's Clarification Form

4 Baseline & Experimental Runs

We used the Lemur IR toolkit (http://www-2.cs.cmu.edu/~lemur/), to conduct our retrieval experiments, with its basic defaults for indexing, and TFIDF for retrieval. We made use of a basic stop word and acronym list, but we did not use a stemmer. Although we used the standard document corpus provided by the LDC, we were unable to successfully remove the foreign language and empty documents that were reported to the track for our experiments. Our baseline run consisted of using the text from the *title* and *description* fields. We used this information for our baseline run because we felt that it most closely approximated the length of queries typically posed by searchers in online searching environments [7]. Indeed, using text from both of these fields created queries that were longer than what is expected, but we did not want our baseline run to produce particularly poor results either.

Our experimental runs were constructed from the information that searchers provided in the *topic narrative* when generating the initial topic descriptions, and from the information that we obtained from searchers with our clarification form (Q2, Q3 and Q4). These various runs are displayed in Table 1. We did not use any metadata in our experimental runs. As described above, we selected the text from the *title* and *description* fields to use for our baseline run. We decided to exclude the data provided in the *topic narrative* field because the quality and amount of text provided in the *topic narrative* field by searchers substantially differed from the text of typical search queries. Thus, we viewed the information provided in the topic narrative as additional information that would likely be provided by searchers' via some post-query elicitation technique, rather than at the time of initial querying, and used it as a source of terms for our experimental runs.

We conducted two major types of experimental runs: automatic and manual. We used query expansion for each of these. Automatic runs included all of the terms that searchers provided in the topic narrative field or in response to Q2, Q3, or Q4. For the manual runs, we (the three authors) each independently examined the contents of searchers' responses and extracted a list of terms that we thought would be potentially useful for query expansion. For three of the sources of terms (topic narrative, Q2, and Q3), our manual extraction resulted in three lists of terms for each topic (3 * 3 * 45). To arrive at a final list of terms for each experimental run, we took the union of the three lists for each topic. Our definition of union was not strict; as long as two people listed a term, it was included in the experimental run. We included these manual runs because we were interested in selecting the most useful terms to use for query expansion in the experimental run rather than all of the terms, and we were not particularly confident that we had the tools to do this automatically. We were further interested in comparing the performance of these two types of runs to investigate the implications of quantity of terms versus quality of terms for retrieval. In other words, are more terms always better, or is the quality of the terms equally, or even more important? The goal of independently extracting these terms and using the union of our lists was to introduce some reliability into the selection of these terms.

We also included an experimental run using the data that we elicited with Q4 of our clarification form. For this question, we only included an automatic experimental run which included the entirety of searchers' responses. Finally, we conducted several experimental runs which consisted of various combinations of the other single-item experimental runs.

Table 1. Experimental runs

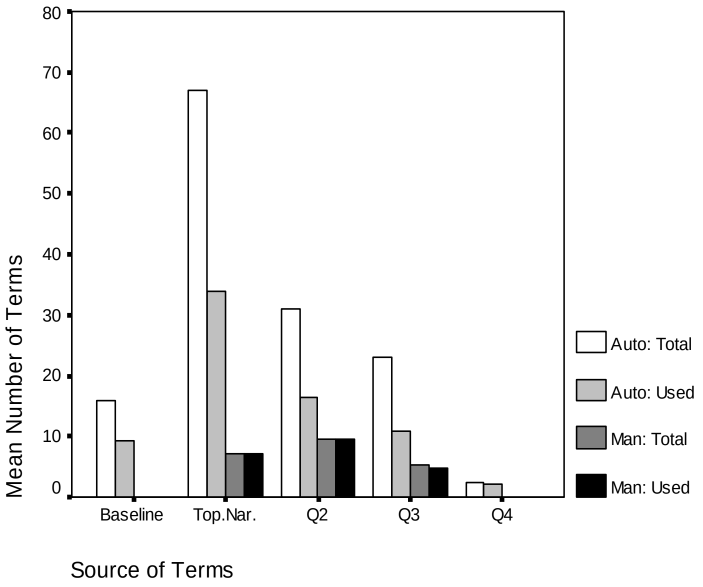
		Technique for Extracting Terms	
		Automatic	**Manual**
Source of Terms	**Topic Narrative**	tn	tnm
	Clarification Form Q2	q2	q2m
	Clarification Form Q3	q3	q3m
	Clarification Form Q4	q4	-
	Combination	q3q4	q3q4m
		q2q3	q2q3m
		q2q4	q2q4m
		q2q3q4	q2q3q4m

5 Results and Discussion

In this section, we first present the descriptive results of our various experimental techniques. This includes results from the topic narrative, and Q2, Q3, Q4 from the clarification form. This is followed by a presentation and comparison of retrieval results for each technique. We conclude this section with a presentation of the descriptive results from our alternative familiarity measure. This section includes an extended exploration of familiarity as it relates to the results of our experimental elicitation techniques. Unless otherwise noted, topics for which no relevant documents were returned (401, 403, 433, 438 and 450) are excluded from analysis. It should also be noted that for both automatic and manual runs we did not remove duplicate terms, but rather used multiple occurrences of a term for weighting. For instance, if a term appeared in a searcher's baseline query and in their response to Q2, then it was counted twice and weighted accordingly.

5.1 Experimental Techniques

A description of the number of terms searchers provided in their topic narratives and in their responses to Q2, Q3, and Q4 is displayed in Figure 2. This Figure presents data that describes the results of both our automatic and manual techniques, and data that describes each technique according to the total number of terms provided, as well as the total number of terms used in retrieval. The total number of terms provided (*total*) was a raw count of the number of terms in each response, while the total number used in retrieval (*used*) was the actual number of terms that were used in retrieval. All results reported below represent total terms and not unique terms. As a reminder, Q2 asked searchers to describe what they already know about the topic, Q3 asked searchers why they wanted to know about the topic, and Q4 asked searchers to input additional keywords describing their topics.



Source	Automatic		Manual	
	Total	Used	Total	Used
Baseline	15.93 (6.01)	9.33 (4.38)	-	-
Top. Narrative	66.91 (31.17)	33.87 (15.51)	7.09 (3.01)	7.07 (3.04)
Q2	30.98 (22.78)	16.53 (11.74)	9.60 (6.47)	9.56 (6.46)
Q3	23.11 (16.45)	10.73 (7.16)	5.42 (3.95)	4.69 (3.53)
Q4	2.47 (4.33)	2.11 (2.93)	-	-

Figure 2. Results of experimental techniques: Mean number of terms (standard deviation)

The average length of searchers' responses to Q2 of the clarification form was 30.98 terms. Of these terms, an average of 16.53 terms were used in retrieval. From the standard deviation it is obvious that the length of searchers' response to this question greatly varied. However, all but five searchers' provided some type of response to this question. Our selection of terms from these responses for inclusion in our manual runs resulted in the selection of, on average, 9.60 terms per topic. Of these terms, 9.56 were used in retrieval. All but three searchers responded to Q3. The average length of these responses was 23.11 terms. Of these terms, an average of 10.73 was used in retrieval. 5.42 were manually selected, and 4.69 were used for retrieval in the manual technique. Finally, only about half of the searchers responded to Q4.

On average, these 23 searchers provided 2.47 terms, of which an average of 2.11 terms were used in retrieval. Interestingly, we noted a large number of spelling errors in searchers' responses to Q2 and Q3 (18 and 10, respectively), which we corrected. It is also interesting that there is little difference in the total number of manually selected terms and the actual number used in retrieval for the topic narrative and for Q2. This is in contrast to Q3, where the difference in these numbers is a bit more.

We were a bit surprised by the results of Q4, by both the actual number of searchers who responded and by the average number of terms that these searchers provided. Given the success of some groups from last year's experiments using a similar question, we expected to elicit more terms with this question than with Q2 or Q3. This was especially true since we expected searchers to be primed to respond to this question based on their interactions with clarification forms provided by other sites. The low response rate to Q4 might be a result of searchers' preference for communicating in natural language rather than keywords. These results might further provide some support that Q2 and Q3 from our clarification form are better techniques for eliciting information from searchers about their information problems than Q4. However, these results might also be explained by an order effect. Unfortunately, we did not set up our clarification form to explicitly compare the differences in these techniques. Instead, questions were always presented in the exact same order, rather than rotated. Thus, it might be the case that searchers were just more fatigued by the time they reached Q4, or out of time, or did not feel that they had anything new to add. Therefore, we are limited in what we can say about these results.

5.2 Baseline and Experimental Runs

The results of our official baseline and experimental runs are displayed in Table 2. In this table, runs are sorted according to *r-precision*, with the best scoring run appearing first. The other standard evaluation measures are also included in this table, along with the median score for each measure across all participating sites. Finally, this table contains information on the average number of new terms used in retrieval for each experimental technique (*average terms added*), along with *average query length*. The original query length in all cases is equal to the baseline (9.3); thus, *average query length* is the sum of the original query length (9.3) and *average terms added*. Because none of our experimental runs made use of metadata, the results in Table 2 are based on the soft+hard scores. However, we note that there was little difference between our hard+soft and hard-only scores.

As can be seen from the table, overall, our official retrieval results are quite poor; in no case did we even approach the median scores for any measure. Performance increased for almost all of our experimental runs, even though in most cases this increase was negligible (less than .0135). Obviously, we suspect that something was not quite right with how we used Lemur[1]. Thus, all analyses and interpretations reported in this section are very tenuous.

Our top performing techniques were both manual, and involved the use of more than one type of clarification form data. The top four techniques were q2q3q4 and q2q4, both the automatic and manual. The performance of q2 (both automatic and manual) from the clarification form followed these. Interestingly, the use of the topic narrative data (both manual and automatic) resulted in performance below our baseline, as did the use of q3. There is at least a consistent pattern to our results, with each technique clustered and ordered in a somewhat systematic way. Many of the techniques derived from the clarification form data performed best when in combination with one another rather than in isolation. For instance, q3 is the worst performing technique in isolation, but when combined with q2 and q4 provides the best performance. Indeed, the addition of q3 adds something to q2 and q4 that they cannot achieve alone or in combination with one another, even though q2 and q4 in isolation performed better than any single item technique. We suspect that searchers' responses to q3, which asked them why they were interested in a particular topic contained many terms that were not useful for defining the topic, but perhaps useful for clarifying it, while responses to q2 and q4, which asked for descriptions of their topics contained more terms that were useful in defining the topic.

We conducted correlations to explore the relationship between query length and performance. There was no statistically significant correlation between query length and any of the three performance

[1] After we submitted our official runs, we discovered that we had a problem with our index. We rebuilt the index, obtained much better results, and were able to see very large differences in retrieval performance according to our experimental techniques, although there was still virtually no difference between the automatic and manual runs. In this paper, we only report and discuss our official TREC results, but are working to publish the other results.

measures. Given that performance was virtually identical across all retrieval techniques, this result is not surprising. We examined the scatterplots for each performance measure and query length to see if perhaps performance increased up to a certain point and then dropped. Our examination of the scatterplots revealed no such relationship.

Table 2. Baseline and experimental results

RUN_ID	Average Terms Added	Average Query Length	R Precision	Average Precision	Precision @ 10
HARD.doc.soft+hard (Median)	**--**	**--**	**0.2906**	**0.2634**	**0.398**
q2q3q4m.eval	17.4	26.7	0.1767	0.1565	0.231
q2q4m.eval	11.9	21.3	0.1761	0.1549	0.247
q2q3q4.eval	29.3	38.6	0.1751	0.1532	0.236
q2q4.eval	18.6	28.0	0.175	0.1539	0.238
q2.eval	16.3	25.5	0.1739	0.1521	0.24
q2m.eval	9.6	19.0	0.1737	0.1521	0.244
q4.eval	2.3	11.7	0.1729	0.1474	0.236
q2q3m.eval	15.1	24.4	0.1655	0.1503	0.233
q3q4.eval	13.0	22.3	0.1647	0.1441	0.222
q2q3.eval	27.0	36.3	0.1635	0.1481	0.236
q3q4m.eval	7.8	17.1	0.1635	0.1424	0.218
BASELINE	**--**	**9.3**	**0.1632**	**0.143**	**0.231**
tn.eval	35.4	44.8	0.1602	0.1478	0.222
tnm.eval	7.3	16.7	0.1602	0.1446	0.229
q3m.eval	5.5	14.8	0.1554	0.1395	0.213
q3.eval	10.6	20.0	0.1536	0.1397	0.22

Finally, we examined the relationship between our best performing manual and automatic techniques to see if the manual techniques, which clearly are more costly in terms of human effort and time, yielded better results than their automatic counterparts. Results according to all performance measures indicated that there is little difference in these two types of techniques. These results suggest that there was little value added during the manual extraction of terms, and that the quality of terms, at least as far as *this* retrieval system is concerned in *this* experiment, may not be reliably determined by humans. However, note that for the top four results, the manual techniques did out-performed their automatic counterparts, albeit by a small margin. These results suggest that perhaps some value is gained in the manual selection of terms. Although query lengths for the automatic techniques were longer than for their manual counterparts, the manual techniques outperformed the automatic. Again, we caution our readers since the differences in performance are so minimal and our overall results suggest that some error likely occurred with our use of the system.

5.3 Familiarity

We included one familiarity measure as part of our clarification form (Q1). As mentioned above, while we did not explicitly collect this information for use in our experiment we thought that the experimental setup of the HARD track provided a nice opportunity to investigate an alternative method for measuring familiarity. The results of our familiarity measure, along with the results of the familiarity measure included in this year's HARD metadata are displayed in Table 3. These results include all searchers (n=50).

Table 3. Familiarity measures

		UNC CF Familiarity: How many times have you searched for information about this topic in the past?				
		Never	1 or 2 Times	3 or 4 Times	5 or more Times	TOTAL
HARD Metadata Familiarity	little	7	17	1	4	29
	much	0	7	4	10	21
	TOTAL	7	24	5	14	50

The results demonstrate that 83% of searchers who indicated that they had searched for information about the topic two or fewer times in the past, also indicated that they had little knowledge of a topic. Accordingly, 67% of those who indicated that they had searched for information three of more times in the past also indicated that they had much knowledge of a topic. For the most part, these results are what one might expect: those who indicated little knowledge of a topic indicated that they had searched fewer times than those who indicated much knowledge of a topic. A Chi-square test provided statistical support for this claim, $\chi^2(3)=14.63$, *p*=.001. These results suggest that one might be able to estimate searchers' familiarity with a topic by tracking their search behaviors over time, and even update this estimate as searchers' continue to search more and more.

To better understand familiarity and its relationship to searchers' behaviors, we conducted several t-tests to see if the length of searchers' title+descriptions (baseline), topic narratives, responses to Q2 and Q3 differed according to their knowledge of the topic. For these analyses, we used the familiarity measure that was part of the HARD metadata and only those topics for which relevant documents were retrieved (n=45). We looked at four measures of length for each item: (1) the total number of terms (i.e. what we used in our automatic run) and (2) the total number used for retrieval; (3) the total number of selected terms (i.e. what we used in our manual run) and (4) the total number of these terms used in retrieval. These means are displayed in Table 4. Results of t-tests demonstrated that there were significantly more selected terms in the topic narratives of searchers' with much knowledge of a topic than those with little knowledge of a topic, *t*(43)=-2.78, *p*<.00. There were no statistically significant differences between familiarity level according to any of the measures involving Q2 or Q3, although searchers with more familiarity with a topic consistently entered longer responses to Q2 and Q3 than did those searchers with little familiarity. Together, these results provide some weak evidence for the notion that a searcher's familiarity level with a topic might be inferred by examining the language that is used to describe the topic.

Table 4. Mean number of terms according to familiarity level; (standard deviation); ns=not significant

				Familiarity		
				Little	**Much**	**Sig.**
Source of Terms	**Baseline**	-	Total	16.58 (6.38)	15.05 (5.52)	ns
		-	Used	9.62 (4.73)	9.05 (3.84)	ns
	Topic Narrative	Automatic	Total	66.65 (23.52)	67.26 (40.09)	ns
			Used	33.31 (12.28)	34.63 (19.44)	ns
		Manual	Total	6.08 (2.29)	8.47 (3.49)	*p*<.00
			Used	6.23 (2.47)	8.21 (3.43)	ns
	CFQ2	Automatic	Total	29.62 (23.25)	32.84 (22.63)	ns
			Used	15.88 (12.82)	17.42 (10.34)	ns
		Manual	Total	9.23 (6.79)	10.11 (6.14)	ns
			Used	9.23 (6.79)	10.00 (6.12)	ns
	CFQ3	Automatic	Total	21.23 (13.16)	25.68 (20.22)	ns
			Used	10.15 (6.32)	11.53 (8.29)	ns
		Manual	Total	4.35 (3.48)	5.16 (3.64)	ns
			Used	5.27 (3.87)	5.63 (4.15)	ns

6 Conclusions

It is hard to make any conclusions since overall, our retrieval results were very poor. We found large differences in the length of searchers' responses to each of the elicitation questions that we used on our clarification forms. Although we are unable to state that one question is better than the other with respect to how much new and useful information is elicited from searchers, we were excited to see that searchers were willing to provide such lengthy responses to some of our questions. One of our original motivations for this experiment was to identify a follow-up technique that could be used in conjunction with some measure of query goodness such as the query clarity measure. Again, our techniques were successful at eliciting more information from searchers; if this information actual improves a query's clarity score and consequently, its retrieval performance, is yet to be determined. After we submitted our official results, we discovered a problem with our index. We rebuilt our index and obtained better and more interesting retrieval results. In doing this, we are in a much better position to more critically and confidently investigate the relationship between query length and retrieval performance.

Finally, we found some interesting results with respect to familiarity. In particular, our measure of familiarity, which is one that might conceivably be measured by examining a person's online search history, was positively associated with the measure used in this experiment. Further, we found very weak evidence that familiarity might be inferred by examining the language that searchers use to describe their topic. Given the recent interest in familiarity, it seems clear that arriving at a valid and reliable estimate of this attribute is an important topic for future research.

Acknowledgements

We would like to thank David Harper for sharing his time and expertise with us.

References

[1] Allan, J. (2003). Hard Track overview in TREC 2003 high accuracy retrieval from documents. In E. Voorhees & L. P. Buckland (Eds.), *TREC-2003, Proceedings of the Twelfth Text Retrieval Conference*. Washington, D. C.: Government Printing Office.

[2] Belkin, N. J., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Query length in interactive information retrieval. In *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03),* Toronto, CA, 205-212.

[3] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR '02), Tampere, Finland, 299-306.

[4] Ingwersen, P. (1982). Search procedures in the library analyzed from the cognitive point of view. *Journal of Documentation, 38*, 165-191.

[5] Kalgren, J. & Franzen, K. (1997). *Verbosity and interface design*. Retrieved on 08 October 2004 at http://www.ling.su.se/staff/franzen/irinterface.html.

[6] Katz, W. A. (2002). *Introduction to reference work: reference services and reference processes, volume 2 (8th Edition).* NY: McGraw-Hill.

[7] Spink, A. & Jansen, B. J. (2004). *Web search: Public searching of the web*. The Netherlands: Kluwer Academic Publishers.