Evaluating Semantic Search Query Approaches with Expert and Casual Users*

Khadija Elbedweihy, Stuart N. Wrigley, and Fabio Ciravegna

Department of Computer Science, University of Sheffield, UK
{k.elbedweihy, s.wrigley, f.ciravegna}@dcs.shef.ac.uk

Abstract. Usability and user satisfaction are of paramount importance when designing interactive software solutions. Furthermore, the optimal design can be dependent not only on the task but also on the type of user. Evaluations can shed light on these issues; however, very few studies have focused on assessing the usability of semantic search systems. As semantic search becomes mainstream, there is growing need for standardised, comprehensive evaluation frameworks. In this study, we assess the usability and user satisfaction of different semantic search query input approaches (natural language and view-based) from the perspective of different user types (experts and casuals). Contrary to previous studies, we found that casual users preferred the form-based query approach whereas expert users found the graph-based to be the most intuitive. Additionally, the controlled-language model offered the most support for casual users but was perceived as restrictive by experts, thus limiting their ability to express their information needs.

1 Introduction

Semantic Web search engines (e.g. Sindice [1]) offer gateways to locate Semantic Web documents and ontologies; ontology-based natural language interfaces (e.g. NLP-Reduce [2]) and visual query approaches (e.g. Semantic Crystal [2]) allow more user-friendly querying; while others try to provide the same support but on the open Web of Data [3, 4]. These search approaches require and employ different query languages. Free-NL provides high expressiveness by allowing users to input queries using their own terms (keywords or full sentences). Controlled-NL provides support during query formulation through suggestions of valid query terms found in the underlying – restrictive – vocabulary.

Finally, view-based (graphs and forms) approaches aim to provide the most support to users by visualising the search space in order to help them understand the available data and the possible queries that can be formulated.

Evaluation of software systems – including user interfaces – has been acknowledged in literature as a critical necessity [5, 6]. Indeed, large-scale evaluations foster research and development by identifying gaps in current approaches and suggesting areas for improvements and future work. Following the Cranfield

* This work was partially supported by the European Union 7th FWP ICT based e-Infrastructures Project SEALS (Semantic Evaluation at Large Scale, FP7-238975).

model [7] – using a test collection, a set of tasks and relevance judgments – and using standard evaluation measures such as precision and recall has been the dominant approach in IR evaluations, led by TREC [8]. This approach has not been without criticisms [9, 10] and there have been long-standing calls for assessing the interactive aspect as well [11, 12].

In an attempt to address these issues, more studies have been conducted with a focus on Interactive Information Retrieval (IIR). The ones embodied within TREC (*Interactive Track* [13] and *Complex Interactive Question-Answering* [14]) involved real users to create topics or evaluate documents rather than to assess usability and usefulness of the IR systems. Others investigated users perception of ease-of-use and user control with respect to the effectiveness of the retrieval process [15] or studied the impact and use of cross-language retrieval systems [16]. With respect to the type of users involved in these studies, some [17,18] have opted to further differentiate between *casual users* and *expert users*. In the context of these works and indeed in ours, *casual users* refer to those with very little or no knowledge in a specific field (e.g., Semantic Web, for our study), while *expert users* have more knowledge and experience in that field.

Inheriting IR's evaluation paradigm, Semantic Search evaluation efforts have been largely performance-oriented [6, 19] with a limited attention to the user-related aspects [20,21]. Kaufmann and Bernstein [20] conducted a within-subjects (same group of subjects evaluate all the participating tools) evaluation of four tools adopting NL- and graph-based approaches with 48 casual users while the evaluation described in [21] featured NL- and form-based tools.

The evaluation described here is different in the following ways: 1) broader range of query approaches (in contrast to [20, 21]), 2) all tools are evaluated within-subjects (in contrast to [21]), and 3) equal-sized subjects groups for casual and expert users (in contrast to [20, 21]). These differences are important and allow novel analyses to be conducted since it facilitates direct comparison of the evaluated approaches and a first-time understanding and comparison of how the two types of users perceive the usability of these approaches. Although some IIR studies involved casual and expert users, most of these focused on investigating differences in the search behaviour and strategies [17, 18, 22].

The remainder of the paper is organized as follows: first, the usability study is described. Next, the results and analyses are discussed together with the main conclusions and finally, the limitations are pointed out with planned future work.

2 Usability Study

The underlying question of the research presented in this paper is how users perceive the usability of different semantic search approaches (specifically support in query formulation and suitability of results returned), and whether this perception is different between expert and casual users. To answer the question, ten casual users and ten expert users were asked to perform five search tasks with five tools adopting NL-based and view-based query approaches. These are user-centric semantic search tools (e.g. query given as natural language or using a form or a graph) querying a repository of semantic data and returning answers

extracted from them. The results returned must be answers rather than documents; however they are not limited to a specific style (e.g., list of entity URIs or visualised results). Experiment results such as query input time, success rates and input of questionnaires are recorded. These results are quantitatively and qualitatively analysed to assess tools' usability and user satisfaction.

2.1 Dataset and Questions

The main requirement for the dataset is to be from a simple and understandable domain for users to be able to formulate the given questions into the tools' query languages. Hence, the geography dataset within the Mooney Natural Language Learning Data[1] was selected. It contained predefined English language questions and has been used by other related studies [20,23]. The five evaluation questions (given below) were chosen to range from simple to complex ones and to test tools' ability in supporting specific features such as comparison or negation.

1. *Give me all the capitals of the USA?*
   This is the simplest question: consisting of only one ontology concept: '*capital*' and one relation between this concept and the given instance: *USA*.
2. *What are the cities in states through which the Mississippi runs?*
   This question contains two concepts: '*city*' and '*state*' and two relations: one between the two concepts and one linking *state* with *Mississippi*.
3. *Which states have a city named Columbia with a city population over 50,000?*
   This question features comparison for a datatype property *city population* and a specific value (50,000).
4. *Which lakes are in the state with the highest point?*
   This question tests the ability for supporting superlatives (*highest point*).
5. *Tell me which rivers do not traverse the state with the capital Nashville?*
   Negation is a traditionally challenging feature for semantic search [24, 25].

2.2 Experiment Setup

Twenty subjects were recruited for the evaluation; ten of these subjects were *casual users* and ten were *expert users*. The 20 subjects (12 females, 8 males) were aged between 19–46 with a mean of 30 years. The experiment followed a within-subjects design to allow direct comparison between the evaluated query approaches. Additionally, with this design, usually less participants are required to get statistically significant results [26]. All 20 subjects evaluated the five tools in randomised order to avoid any learning, tiredness or frustration effects that could influence the experiment results. Furthermore, to avoid any possible bias introduced by developers evaluating their own tools, only one test leader – who is also not the developer of any of the tools – was responsible for running the whole experiment.

For each tool, subjects were given a short demo session explaining how to use it to formulate queries. After that, subjects were asked to formulate each of

[1] `http://www.cs.utexas.edu/users/ml/nldata.html`

the five questions in turn using the tool's interface. The order of the questions was randomised for each tool to avoid any learning effects. After testing each tool, subjects were asked to fill in two questionnaires.

Finally, we collected demographics data such as age, profession and knowledge of linguistics (see [27] for details of all three questionnaires). Each experiment with one user took between 60 to 90 minutes.

In assessing usability of user-interfaces, several measurements including time required to perform tasks, success rate and perceived user satisfaction were proposed in the literature of IIR [28, 29] and HCI [30, 31].

Similar to these studies and indeed to allow for deeper analysis, we collected both objective and subjective data covering the experiment results. The first included: 1) *input time* required by users to formulate their queries, 2) *number of attempts* showing how many times on average users reformulated their query to obtain answers with which they were satisfied (or indicated that they were confident a suitable answer could not be found), and 3) *answer found rate* capturing the distinction between finding the appropriate answer and the user 'giving up' after a number of attempts. This data was collected using custom-written software which allowed each experiment run to be orchestrated.

Additionally, subjective data was collected using think-aloud strategy [32] and two post-search questionnaires. The first is the *System Usability Scale (SUS) questionnaire* [33], a standardised usability test consisting of ten normalised questions covering aspects such as the need for support, training, and complexity and has proven to be very useful when investigating interface usability [34]. The second questionnaire (*Extended Questionnaire*) is one which we designed to capture further aspects such as the user's satisfaction with respect to the tool's query language and the content returned in the results as well as how it was presented. After completing the experiment, subjects were asked to rank the tools according to four different criteria (each one separately): how much they liked the tools (*Tool Rank*); how much they liked their query interfaces: graph-based, form-based, free-NL and controlled-NL (*Query Interface Rank*); how much they found the results to be informative and sufficient (*Results Content Rank*); and finally how much they liked the results presentation (*Results Presentation Rank*). Note that users were allowed to give equal rankings for multiple tools if they had no preference for one over the other. To facilitate comparison, for each criterion, ranking given by all users for one tool was summed and subsequent score was then normalised to have ranges between 0 and 1 (where 1 is the highest).

3 Results and Discussion

Evaluated tools included free-NL- (NLP-Reduce [2]), controlled-NL- (Ginseng [2]), form- (K-Search [35]), and finally graph- based (Semantic-Crystal [2] and Affective Graphs[2]) approaches. Results for both expert and casual users are presented in Tables 1 and 2 respectively. In these tables, a number of different factors are reported such as the SUS scores and the tools' rankings. We also include the

[2] `http://oak.dcs.shef.ac.uk/?q=node/253`

Table 1. Tools results for expert users. Non-ranked scores are median values; bold values indicate best performing tool in that category.

Criterion	Affective Graphs	Semantic Crystal	K-Search	Ginseng	NLP-Reduce	p-value
SUS (0-100)	**63.75**	50	40	32.5	37.5	0.003
Tool Rank (0-1)	**0.875**	0.625	0.6	0.225	0.225	-
Query Language Rank (0-1)	**0.925**	0.725	0.65	0.425	0.45	-
Results Content Rank (0-1)	0.875	0.875	**0.925**	0.725	0.725	-
Results Presentation Rank (0-1)	0.875	0.875	**0.975**	0.8	0.8	-
EQ1: liked presentation (0-5)	2.5	2.5	**4**	3	3	0.007
EQ2: query language easy (0-5)	**4**	**4**	**4**	2	2.5	0.035
Number of Attempts	**1.5**	2.2	2	1.7	4.1	0.001
Answer Found Rate (0-1)	**0.8**	0.4	0.5	0.4	0.2	0.004
Input Time (s)	88.86	79.55	53.54	102.52	**19.90**	0.001

scores from two of the most relevant questions from the extended questionnaire. *EQ1: liked presentation* shows the average response to the question "I liked the presentation of the answers", while *EQ2: query language easy* shows it for the question "The system's query language was easy to use and understand".

Note that in the rest of this section, we use the term *tool* (e.g. graph-based tools) to refer to the implemented tool as a full semantic search system (with respect to its query interface and approach, functionalities, results presentation, etc.) and the term *query approach* (e.g. graph-based query approach) to specifically refer to the style of query input adopted.

To quantitatively analyse the results, SPSS[3] was used to produce averages, perform correlation analysis and check the statistical significance. The median (as opposed to the mean) was used throughout the analysis since it was found to be less susceptible to outliers or extreme values sometimes found in the data. In the qualitative analysis, the open coding technique [36] was used in which the data was categorised and labelled according to several aspects dominated by usability of the tools' query approaches and returned answers.

3.1 Expert User Results

According to the adjective ratings introduced by [37], Ginseng – with the lowest SUS score – is classified as *Poor*, NLP-Reduce as *Poor* to *OK*, K-Search and Semantic Crystal are both classified as *OK*, while Affective Graphs, which managed to get the highest average SUS score, is classified as *Good*. These results are also confirmed by the tools' ranks (see Table 1): Affective Graphs was selected 60% of the times as the most-liked tool and thus got the highest rank (0.875), followed by Semantic Crystal and K-Search (0.625 and 0.6 respectively) and finally Ginseng and NLP-Reduce got a very low rank (0.225) with each being chosen as the least-liked tools four times and twice, respectively. Since the rankings are an inherently relative measure, they allow for direct tool-to-tool comparisons to be made. Such comparisons using the SUS questionnaire may be less reliable since the questionnaire is completed after each tool's experiment (and thus temporally spaced) with no direct frame of reference to any of the other tools.

Table 1 also shows that Affective Graphs, which is most liked and found to be the most intuitive by users managed to get satisfactory answers for 80% of the

[3] `www.ibm.com/software/uk/analytics/spss/`

queries, followed by K-Search (50%) which is employing the second most-liked query approach. Finally, it was found that all the participating tools did not support negation (except partially by Affective Graphs). This was confirmed by the *answer found rate* for the question "*Tell me which rivers do **not** traverse the state with the capital nashville?*" being: Affective Graphs: 0.4, Semantic Crystal: 0.1, K-Search: 0.1, Ginseng: 0.1, NLP-Reduce: 0.0.

Expert users prefer graph- and form- based approaches: Results showed that graph- and form- based approaches were the most liked by expert users. However, in terms of overall satisfaction (see SUS scores and Tool Rank in Table 1), graph-based tools outperformed the form- and NL- based ones. Additionally, feedback showed that *users were able to formulate more complex queries with the view-based approaches (graphs and forms) than with the NL ones (free and controlled)*. Indeed, the ability to visualise the search space provides an understanding of the available data (concepts) as well as connections found between them (relations) which shows how they can be used together in a query [20, 38].

It is interesting to note that although Affective Graphs and Semantic Crystal both employ graph-based query approach, users had different perceptions of their usability. More users gave the query interface of Affective Graphs higher scores than Semantic Crystal (quartiles: "3.75 , 5" and "2 , 4.25" respectively) since they found it to be more intuitive. The most repeated (60%) *positive* comment given for Affective Graphs was "*the query interface is intuitive and easy/pleasant to use*". This is a surprising outcome since graph-based approaches are known to be complicated and laborious [20, 38]. However, this has not been explicitly assessed from expert users perspective in any similar studies.

An important difference was observed between the two graph-based tools: Semantic Crystal visualizing the entire ontology whereas Affective Graphs opted for showing concepts and relations only selected by the users (see Fig. 1). Although feedback showed that users preferred the first approach, it imposes a limitation on how much can be displayed in the visualisation window. With a small ontology, the graph is clear and can be easily explored; as the ontology gets bigger, the view would easily get cluttered with concepts and links showing relations between them. This would negatively affect the usability of the interface and in turn the user experience.

Expert users frustrated by controlled-NL: Although the guidance provided by the controlled-NL approach was at sometimes appreciated, restricting expert users to the tool's vocabulary was more annoying. This resulted in an unsatisfying experience (lowest SUS score of 32.5 and least liked interface) which is supported by the most repeated *negative* comments given for Ginseng:

- It is frustrating when you cannot construct queries in the way you want.
- You need to know in advance the vocabulary to be able to use the system.

The second comment is in stark contrast to what the controlled-NL approach is designed to provide. It is intended to help users formulate their queries without having to know the underlying vocabulary. However, even with the guidance, users frequently got stuck because they did not know how to associate the

Table 2. Tools results for casual users. Non-ranked scores are median values; bold values indicate best performing tool in that category.

Criterion	Affective Graphs	Semantic Crystal	K-Search	Ginseng	NLP-Reduce	p-value
SUS (0-100)	55	**61.25**	41.25	53.75	43.75	0.485
Tool Rank (0-1)	**0.675**	**0.675**	0.575	0.45	0.275	-
Query Language Rank (0-1)	0.525	0.55	**0.625**	0.525	0.4	-
Results Content Rank (0-1)	0.675	0.75	**0.775**	0.575	0.575	-
Results Presentation Rank (0-1)	0.775	0.7	**0.8**	0.6	0.475	-
EQ1: liked presentation (0-5)	3	3	**3.5**	2.5	2	0.3
EQ2: query language easy (0-5)	**4**	**4**	**4**	3	3	0.131
Number of Attempts	**1.7**	1.8	2.1	**1.7**	4.2	0.001
Answer Found Rate (0-1)	0.4	**0.6**	0.5	0.4	0.2	0.150
Input Time (s)	72.8	75.76	63.59	93.13	**18.6**	0.001

suggested concepts, relations or instances together. This is confirmed by users requiring the longest input time when using Ginseng (Table 1 : Input Time).

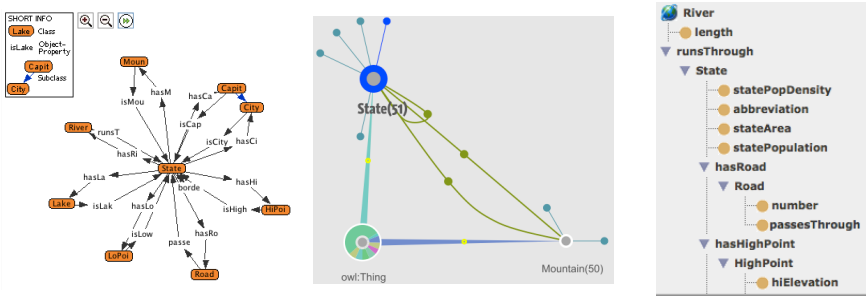
3.2 Casual User Results

Graph-based tools more complex if entire ontology not shown: Recall in Section 3.1, expert users preferred the approach of visualising the entire ontology (adopted by Semantic Crystal as shown in Fig. 1a). This was indeed more appreciated by casual users, resulting in Semantic Crystal receiving higher scores. Surprisingly, the lack of this feature caused Affective Graphs to be perceived by casual users as the most complex and difficult to use: 50% of the users found it to be: "*less intuitive and has higher learning curve than NL*".

Tool interface aesthetics important to casual users: Most of the casual users (70%) liked the interface of Affective Graphs for having an *animated, modern and visually-appealing* design. This not only created a pleasant search experience but was also helpful during query formulation (e.g., highlighting selected concepts) and in turn balanced the negative effect of not showing the entire ontology, resulting in high user satisfaction (second highest SUS score: 55).

Casual users prefer form-based approach: Casual users needed less input time with the form-based approach and found it less complicated than the graph-based approach while allowing more complex queries than the NL-based ones. However, unexpectedly, more attempts were required to formulate their queries using this approach. The presence of inverse relations in the ontology was viewed by casual users as unnecessary redundancy. This impression led to confusion and thus required more trials to formulate the right queries. For instance, to query for the rivers running through a certain state, two alternatives ("State, hasRiver, River" and "River, runsthrough, State") were adopted by users. Tools ought to take the burden off users and provide one unique way to formulate a single query.

Casual users liked controlled-NL support: Casual users found the guidance offered by suggesting valid query terms very helpful and provided them with more confidence in their queries. Interestingly, they preferred to be 'controlled' by the language model (allowing only valid queries) rather than having more expressiveness (provided by free-NL) while creating more invalid queries.



(a) Semantic Crystal (b) Affective Graphs (c) K-Search

Fig. 1. Different visualizations of the Mooney ontology by the tools

3.3 Results independent of user type

This section discusses results and findings common to both types of users.

Form-based faster but more tedious than graph-based: Results showed that both types of users took less time to formulate their queries with the form-based approach than with the graph-based ones (approximate difference: 36% for experts, 14% for casuals). However, it was found to be more laborious to use than graphs especially when users had to inspect the concepts and properties (presented in a tree-like structure) to select the required ones for the query (see Fig. 1c). This is a challenge acknowledged in the literature [39] for form-based approaches and is supported by the feedback given by users: the most repeated negative comment was "*It was hard to find what I was looking for once a number of items in the tree are expanded*". Additionally, this outcome suggests that input time cannot be used as the sole metric to inform usability of query approaches.

Free-NL simplest and most natural; suffer from habitability problem: The free-NL approach was appreciated by users for being the most simple and natural to them. However, the results showed a frequent mismatch between users' query terms and the ones expected by the tool. This is caused by the abstraction of the search space and is known in literature as the *habitability problem* [2, p.2]. This is supported by the users' most repeated negative comment: "*I have to guess the right words*". They found that they could get answers with specific query terms rather than others. For instance, using 'run through' with 'river' returns answers which are not given when using 'traverse'. This is also confirmed by the tool (NLP-Reduce) getting the lowest success rate (20%). Furthermore, requiring the highest *number of attempts* (4.1) support users' feedback that they had to rephrase their queries to find the combination of words the tool is expecting. Indeed, this is a general challenge facing natural language interfaces [20, 40, 41].

Results content and presentation affected usability and satisfaction: When evaluating semantic search tools, it is important – besides evaluating performance and usability – to assess the usefulness of the information returned as well as how it is presented. Within this context, our study found that the

results presentation style employed by K-Search was the most liked by all users as shown in Tables 1 and 2. It is interesting to note how small details such as organising answers in a table or having a visually-appealing display (adopted by K-Search) have a direct impact on results readability and clarity and, in turn, user satisfaction. This is shown from the most repeated comments given for K-Search: "*I liked the way answers are displayed*" and "*results presentation was easy to interpret*". Additionally, K-Search is the only tool that did not present a URI for an answer but used a reference to the document using a NL label. This was favoured by users who often found URIs to be technical and more targeted towards domain experts. For instance, one user specifically mentioned having "*http://www.mooney.net/geo#tennesse2*" as an answer was not understandable. By examining the ontology, this was found to be the URI of *tennessee river* and it had the '2' at the end to differentiate it from *tennessee state*, which had the URI "*http://www.mooney.net/geo#tennesse*". This suggests that, unless users are very familiar with the data, presenting URIs alone is not very helpful. By analysing users feedback from a similar usability study, Elbedweihy et al. [21] found that when returning answers to users, each result should be augmented with associated information to provide a 'richer' user experience. This was similarly shown by users' feedback in our study with the following comments regarding potential improvements often given for all the tools:

- Maybe a 'mouse over' function with the results that show more information.
- Perhaps related information with the results.
- Providing similar searches would have been helpful.

For example, for a query requiring information about states, tools could go a step further and return extra information about each state – rather than only providing name and URI – such as the *capital, area, population* or *density*, among others. Furthermore, they could augment the results with ones associated with related concepts which might be of interest to users [42, 43]. Again, these could be instances of *lakes or mountains* (examples of concepts related to *state*) found in a state. This notion of relatedness or relevancy is clearly domain-dependent and is itself a research challenge. In this context, Elbedweihy et al. [44] suggested a notion of relatedness based on collaborative knowledge found in query logs.

Benefit of displaying generated formal query depends on user type: While casual users often perceived the formal query generated by a tool as confusing, experts liked the ability to see the formal representation of their constructed query since it increased their confidence in what they were doing. Indeed, being able to perform direct changes to the formal query increased the expressiveness of the query language as perceived by expert users.

Experts plan query formulation more than casuals: As shown in Table 3, with most of the tools, expert users took more time to build their queries than casual ones. The feedback showed that the latter often spent more time planning – and verbally describing – their rationale (e.g. "so it understands abbreviations and it seems to work better with sentences than with keywords") during query formulation. Interestingly, studies on user search behaviour found similar results:

Table 3. Query input time (in seconds) required by expert and casual users.

User Type	Affective Graphs	Semantic Crystal	K-Search	Ginseng	NLP-Reduce	p-value
Expert Users	88.86	79.55	53.54	102.52	19.90	0.001
Casual Users	72.8	75.76	63.59	93.13	18.6	0.001

Tabatabai and Shore found that "*Novices were less patient and relied more on trial-and-error.*" [17, p.238] and Navarro-Prieto et al. showed that "*Experienced searchers ... planned in advance more than the novice participants*" [18, p.8].

4 Conclusions

In this paper, we have discussed a usability study of five semantic search tools employing four different query approaches: free-NL, controlled-NL, graph-based and form-based. The study – which used both expert and casual users – has identified a number of findings, the most important are summarised below.

Graph-based approaches were perceived by expert users as intuitive allowing them to formulate more complex queries, while casual users, despite finding them difficult to use, enjoyed the visually-appealing interfaces which created an overall pleasant search experience. Also, showing the entire ontology helped users to understand the data and the possible ways of constructing queries. However, unsurprisingly, graph-based approach was judged as laborious and time consuming. In this context, the form-based approach required less input time. It was also perceived as a midpoint between NL-based and graph-based, allowing more complex queries than the first, yet less complicated than the latter.

Additionally, casual users found the controlled-NL support to be very helpful whereas expert users found it to be very restrictive and thus preferred the flexibility and expressiveness offered by free-NL. A major challenge for the latter was the mismatch between users' query terms and ones expected by the tool (habitability problem). The results also support the literature showing that negation is a challenge for semantic search tools [24, 25]: only one tool provided partial support for negation. Furthermore, the study showed that users often requested the search results to be augmented with more information to have a better understanding of the answers. They also mentioned the need for a more user-friendly results presentation format. In this context, the most liked presentation was that employed by K-Search, providing results in a tabular format that was perceived as clear and visually-appealing.

To conclude, this usability study highlighted the advantage of visualising the search space offered by view-based query approaches. We suggest combining this with a NL-input feature that would balance difficulty and speed of query formulation. Indeed, providing *optional* guidance for the NL input could be the best way to cater to both expert and casual users within the same interface. These findings are important for developers of future query approaches and similar user interfaces who have to cater for different types of users with different preferences and needs. For future work and, indeed, to have a more complete picture, we plan to assess how the interaction with the search tools affect the information seeking process (usefulness). To achieve this, we will use questions with an overall goal – as opposed to ones which are not part of any overarching information need – and compare users' knowledge before and after the search task. This would

also allow us to evaluate advanced features such as formulating complex queries, merging results of subqueries or assessing relevancy and usefulness of additional information presented with the results.

References

1. Tummarello, G., Oren, E., Delbru, R.: Sindice.com: Weaving the Open Linked Data. In: Proc. ISWC/ASWC 2007
2. Kaufmann, E., Bernstein, A.: Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases. J. Web Sem. **8** (2010)
3. López, V., Motta, E., Uren, V.: PowerAqua: Fishing the Semantic Web. In: Proc. ESWC 2006
4. Harth, A.: VisiNav: A system for visual search and navigation on web data. J. Web Sem. **8** (2010) 348–354
5. Saracevic, T.: Evaluation of evaluation in information retrieval. In: Proc. SIGIR 1995
6. Halpin, H., Herzig, D.M., Mika, P., Blanco, R., Pound, J., Thompson, H.S., Tran, D.T.: Evaluating Ad-Hoc Object Retrieval. In: Proc. IWEST 2010 Workshop
7. Cleverdon, C.W.: Report on the first stage of an investigation onto the comparative efficiency of indexing systems. Technical report, The College of Aeronautics, Cranfield, England (1960)
8. Spärck Jones, K.: Further reflections on TREC. Inf. Process. Manage. **36** (2000) 37–85
9. Voorhees, E.: The Philosophy of Information Retrieval Evaluation. In: Proc. CLEF 2001
10. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer (2005)
11. Salton, G.: Evaluation problems in interactive information retrieval. Information Storage and Retrieval **6** (1970) 29 – 44
12. Tague, J., Schultz, R.: Evaluation of the user interface in an information retrieval system: A model. Inf. Process. Manage. (1989) 377–389
13. Hersh, W., Over, P.: SIGIR workshop on interactive retrieval at TREC and beyond. SIGIR Forum **34** (2000)
14. Kelly, D., Lin, J.: Overview of the TREC 2006 ciQA task. SIGIR Forum **41** (2007) 107–116
15. Xie, H.: Supporting ease-of-use and user control: desired features and structure of web-based online IR systems. Inf. Process. Manage. **39** (2003) 899–922
16. Petrelli, D.: On the role of user-centred evaluation in the advancement of interactive information retrieval. Inf. Process. Manage. **44** (2008) 22 – 38
17. Tabatabai, D., Shore, B.M.: How experts and novices search the Web. Library & Information Science Research **27** (2005) 222 – 248
18. Navarro-Prieto, N., Scaife, M., Rogers, Y.: Cognitive strategies in web searching. Proc. the 5th Conference on Human Factors and the Web **2004** (1999) 1–13
19. Balog, K., Serdyukov, P., de Vries, A.: Overview of the TREC 2010 Entity Track. In: TREC 2010 Working Notes
20. Kaufmann, E., Bernstein, A.: How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users? In: Proc. ISWC 2007
21. Elbedweihy, K., Wrigley, S.N., Ciravegna, F., Reinhard, D., Bernstein, A.: Evaluating semantic search systems to identify future directions of research. In: Proc. 2nd International Workshop on Evaluation of Semantic Technologies (IWEST 2012)

22. Hölscher, C., Strube, G.: Web search behavior of Internet experts and newbies. Comput. Netw. **33** (2000) 337–346
23. Popescu, A.M., Etzioni, O., Kautz, H.: Towards a theory of natural language interfaces to databases. In: IUI 2003. (2003) 149–157
24. Damljanovic, D., Agatonovic, M., Cunningham, H.: FREyA: an Interactive Way of Querying Linked Data using Natural Language. In: Proc. QALD-1 Workshop
25. López, V., Fernández, M., Motta, E., Stieler, N.: PowerAqua: supporting users in querying and exploring the semantic web. Semantic Web **3** (2012)
26. Albert, W., Tullis, T., Tedesco, D.: Beyond the Usability Lab: Conducting Large-Scale User Experience Studies. Elsevier Science (2010)
27. Bernstein, A., Reinhard, D., Wrigley, S.N., Ciravegna, F.: Evaluation design and collection of test data for semantic search tools. Technical Report D13.1, SEALS Consortium (2009)
28. Miller, R.: Human Ease of Use Criteria and Their Tradeoffs. IBM, Systems Development Division (1971)
29. Kelly, D.: Methods for Evaluating Interactive Information Retrieval Systems with Users. Found. Trends Inf. Retr. **3** (2009) 1–224
30. Hix, D., Hartson, H.R.: Developing User Interfaces: Ensuring Usability Through Product and Process. J. Wiley (1993)
31. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison Wesley Longman (1998)
32. Ericsson, K.A., Simon, H.A.: Protocol analysis: Verbal reports as data. MIT Press (1993)
33. Brooke, J.: SUS: a quick and dirty usability scale. In: Usability Evaluation in Industry. CRC Press (1996) 189–194
34. Bangor, A., Kortum, P.T., Miller, J.T.: An Empirical Evaluation of the System Usability Scale. Int't J. Human-Computer Interaction **24** (2008) 574–594
35. Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D.: Hybrid Search: Effectively Combining Keywords and Ontology-based Searches. In: Proc. ESWC 2008
36. Strauss, A., Corbin, J.: Basics of qualitative research: grounded theory procedures and techniques. Sage Publications (1990)
37. Bangor, A., Kortum, P.T., Miller, J.T.: Determining what individual SUS scores mean: Adding an adjective rating scale. J. Usability Studies **4** (2009) 114–123
38. Uren, V., Lei, Y., López, V., Liu, H., Motta, E., Giordanino, M.: The usability of semantic search tools: a review. Knowledge Engineering Review **22** (2007) 361–377
39. López, V., Motta, E., Uren, V., Sabou, M.: Literature review and state of the art on Semantic Question Answering (2007)
40. López, V., Uren, V., Sabou, M., Motta, E.: Is question answering fit for the Semantic Web? A survey. Semantic Web **2** (2011) 125–155
41. Uren, V., Lei, Y., López, V., Liu, H., Motta, E., Giordanino, M.: The usability of semantic search tools: a review. The Knowledge Eng. Rev. **22** (2007) 361–377
42. Meij, E., Mika, P., Zaragoza, H.: Investigating the Demand Side of Semantic Search through Query Log Analysis. In: Proc. SemSearch 2009
43. Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Mapping queries to the Linking Open Data cloud: A case study using DBpedia. J. Web Sem. **9** (2011) 418 – 433
44. Elbedweihy, K., Wrigley, S.N., Ciravegna, F.: Improving Semantic Search Using Query Log Analysis. In: Proc. Interacting with Linked Data (ILD) 2012 Workshop