1. Proposal Title
Getting educated into search engines

2. Principal Investigator
Hervé Varenne, Ph.D.
Professor and Director, Center for Everyday Education, Teachers College, Columbia University
Teachers College, Columbia University
525 West 120th Street, New York, New York 10027
212-678-3190 – hhv1@columbia.edu

3. Research Abstract and Goal
Most people who now routinely use search engines learned how to do so informally. This research proposes to identify moments when this informal learning has in some way failed users, such as those who recently did a Google search for "facebook login" and ended up leaving frustrated comments on the blog ReadWriteWeb. This research builds on earlier research by team members that has successful explored new directions for Internet education, filling gaps in the literature on search behavior, and developing user profiles out of "noise" and apparent errors. The methods of this research project point the way to critical elements of natural language use which should strengthen semantic web analyses, helping developers support users better.

4. Keywords
Education innovation; Natural language processing; Human-computer interaction

5. Technical description
Search engines, like all aspects of new computing technologies and media, have emerged late in the life span of most people who now use them routinely, after most people have finished their schooling. Even for younger generations, most education about search engines is not conducted in school, where legal issues have many administrators convinced that Internet access should be tightly controlled or even banned; not to mention that many teachers alive today did not receive formal training in using search engines themselves. And yet learning continues. In the absence of formal training, people learn to use search engines mostly through family and peer networks. Education about search engines is thus a process that has not been tamed by institutions, and is quite open to the production of unintended consequences which may then have serious consequences on the everyday life of individuals, and even nations. From identity theft and information cruft (spam) to resistance to medical expertise (e.g. on matters of vaccination) and scientific expertise (e.g. evolution, or global warming), search engines are implicated in how people inform themselves and make decisions about the world around them.

Some researchers have argued that traditional models in information science research are based on outdated understandings of human behavior; they tend to rely on top-down, "abstract metaphors" to operationalize human behavior (such as the vaguely-defined terms like "information literacy") in formal settings like schools and libraries. Few
models are built from the actual observations of how humans seek, use, and even build knowledge day-to-day from their social networks. (Tuominen, Savolainen, and Talja, 2005) Education and communications studies, through detailed observations of naturally occurring behavior, have shown that people draw on the rich resources they find in their specific, local, social contexts. This work has provided insight into major schisms in user behavior (boyd, 2008) and how individual users teach themselves to use technologies (Ito et al, 2009; Hung, 2009; Gee 2003). These findings suggest a few things: one, that to educate for a wired world, formal education must be more complex and nuanced; and two, that self- and community education are increasingly important and should be better understood.

In software development, as in traditional information science research, understandings of human behavior are often treated in broad quantitative strokes which dismiss outliers as "noise." PageRank provides excellent results for tech-savvy users who are satisfied with pages that have already been linked to, blogged about, tweeted, Slashdotted, or otherwise approved by their peers. But it can put less savvy users on the wrong side of a popularity contest — one their peers can't win because they don't link, write blog posts, tweet, or spend time on BoingBoing or Slashdot, or sometimes even use URLs successfully. One recent example of this phenomenon is the deluge received by ReadWriteWeb, one of Technorati's top twenty blogs, from people doing Google searches for "facebook login" (http://www.readwriteweb.com/archives/web_illiteracy_how_much_is_your_fault.php). Current research by Gillian Andrews of the Center for Everyday Education has documented how much work bloggers and blog commenters do to make sense of the design and limitations of blogging software, drawing on their informal learning and non-Internet literacy skills learned in school. The proposed research builds on this work.

The research methods employed by associates at the Center focus a microscope on outlier "noise," trying to make sense of what may otherwise appear to be nonsensical user behavior. By analyzing the language of conversations in which users puzzle out why a search engine took them to a page, what a page is about, and what they can and cannot do once they are there, this research has been able to identify not only major user misunderstandings, but also ways in which search engine developers, bloggers, and other designers could have better supported their searches.

This grant will support an ongoing research project which explores users' (mis)understandings of search engines (and related Internet-navigation features such as URLs). This project should have two outcomes: one, to provide insight and support for search engine learning both in and out of the classroom, and two, to semantically fine-tune search algorithms to better serve populations who fall through Google's cracks (for example, those using the search terms "cancel google," "please cancel my account," or "facebook login," all of which have led to major misunderstanding pileups on blog comment threads; see for example http://www.jonathancoulton.com/2006/06/13/please-please-cancel-my-account) This research employs sociolinguistic methods which point the way to critical, yet hard-to-pin-down, elements of language (specifically indexicality) which might strengthen semantic web analyses.
The research will proceed as follows:
1) Employ natural language processing methods to identify greater numbers of online conversations resulting from search engine misunderstanding, and analyze them;
   a. Identify collocations and bigrams in online discussions which are indicators of misunderstandings;
   b. Contact and interview users who appear to have trouble navigating search results;
2) Life history interviews with several individuals about their encounters with search engines, including
   a. What resources and people they have worked with to learn more about search engines;
   b. What difficulties they have encountered and how they have been educating themselves on what to do about them;
   c. Drawing maps of their mental models of how search engines and web servers work;
3) Detailed observations of people using search engines to accomplish a joint task (e.g. planning major events like weddings, of college admissions). This requires
   a. Getting to know the immediate network of the people involved, and how they are called upon in seeking information;
   b. Using Morae observation software to record and analyze instances of searching (including setting up the searching, doing it, discussing the results, etc.)

The sample for interviews and observations would include individual families selected from two types: one, more prosperous and experienced users, and two, less prosperous users with less new media experience. While non-random sampling of this sort does not allow probabilistic generalizations to a population, extensive research has shown that it can be extremely helpful in finding possibilities and constraints that large-scale, generalizeable methods cannot reach; it can then suggest new questions for larger-scale surveys.

6. Expected outcomes
1) Recommendations about Internet skills for educators, librarians, and policymakers.
2) Tools to support (at home as well as in more formal settings) comprehension of pages that appear in search results: possibly including a series of online videos or a lightweight browser plugin to support users' understanding of how search engines work, strategies for learning who owns a page, and common errors in URL writing which may redirect users to ISP- or browser-provided search engines.
3) Discussions with search engine developers about the possibility and utility of incorporating findings into search tools.