

2006

A field experimental study of natural language versus Boolean searching

Lisa Ha T Ngo
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Ngo, Lisa Ha T, "A field experimental study of natural language versus Boolean searching" (2006). *Master's Theses*. 2970.
DOI: <https://doi.org/10.31979/etd.hkgj-ray8>
https://scholarworks.sjsu.edu/etd_theses/2970

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

A FIELD EXPERIMENTAL STUDY OF NATURAL LANGUAGE VERSUS
BOOLEAN SEARCHING

A Thesis

Presented to

The Faculty of the School of
Library and Information Science
San José State University

In Partial Fulfillment

of the Requirements for the Degree
Master of Library and Information Science

by

Lisa Ha T. Ngo

August 2006

UMI Number: 1438585

Copyright 2006 by
Ngo, Lisa Ha T.

All rights reserved.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 1438585

Copyright 2007 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

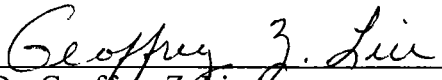
ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© 2006

Lisa Ha T. Ngo

ALL RIGHTS RESERVED


APPROVED FOR THE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE


Dr. Geoffrey Z. Liu


Dr. Ziming Liu


Dr. Jo Bell Whitlatch

APPROVED FOR THE UNIVERSITY


Thea I. Williamson

ABSTRACT

A FIELD EXPERIMENTAL STUDY OF NATURAL LANGUAGE VERSUS BOOLEAN SEARCHING

by Lisa Ha T. Ngo

This field experimental study compares the effectiveness of natural language search with Boolean search with broad queries in an academic environment. Sample queries were collected from undergraduate students with actual information needs at San José State University and searched in a heavily used database available at the university library; the numbers of retrieved citations, retrieved relevant citations, and retrieved unique relevant citations were analyzed. Natural language search retrieved a significantly larger number of total citations for both broad and narrow queries. For broad queries, no significant difference between natural language search and Boolean search was found in either the number of relevant citations or the number of unique relevant citations retrieved. However, for narrow queries, Boolean search was found to be far more successful in terms of the numbers of relevant and unique relevant citations retrieved. Implications of these findings for students, database designers, and librarians are discussed.

ACKNOWLEDGEMENTS

Many, many thanks to my thesis committee - Dr. Geoffrey Liu, Dr. Jo Bell Whitlatch, and Dr. Ziming Liu - for lending their expertise to this project and for their invaluable comments and suggestions. This thesis could not have been completed without your support. My gratitude also goes to the SJSU librarians at the Dr. Martin Luther King, Jr. Library for their help in recruiting the undergraduate participants needed for this study. Lastly, to Caroline, Rosalinda, and all my wonderful friends, thanks for your much needed and always welcome company.

Table of Contents

List of Tables	vii
Introduction.....	1
Literature Review.....	4
Definition of the Research Problem.....	12
Methodology.....	14
Results	24
Discussion	30
Conclusion	37
References.....	39
Appendix A – Relevance Rating Form for Retrieved Citations	41
Appendix B – Information for Students.....	42

List of Tables

Table 1: Query Examples.....23

Table 2: Means of Total Relevant Articles, Total Retrieved Articles, and
Total Unique Relevant Articles Retrieved: Broad Queries25

Table 3: Results of Paired Samples t-Test: Broad Queries.....26

Table 4: Means of Total Relevant Articles, Total Retrieved Articles, and
Total Unique Relevant Articles Retrieved: Narrow Queries27

Table 5: Results of Paired Samples t-Test: Narrow Queries28

Introduction

While the most common searches performed in information retrieval systems (IR) today arguably use Boolean logic, systems that support natural language (NL) search are growing as an increasingly viable alternative for those who don't want to learn the cryptic syntaxes needed for conducting Boolean search. There are two kinds of natural language search that come into play (in varying degrees) in today's commercial systems that claim to support natural language. One is Natural Language Processing (NLP) as an indexing process in IR, specifically defined by Liddy (1997) as "a set of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" (p. 137). The usage of the term "natural language" in NLP is largely confined to researchers and developers of IR systems, and it is hardly ever seen or understood by the average end-user. The other kind is more likely recognizable to database searchers and can be described as "an interface or software system which allows the user of a database system to search for results by asking plain language questions" (Arnold & Rosen, 1993, p. 32). It is important to note that a natural language interface can (and often does) exist independently of natural language processing; indeed, many systems that purportedly support NLP do nothing more than providing an NL interface with shallow, if any, linguistic processing of queries or full texts.

There are, however, several commercial systems that fall in between supporting genuine NLP and merely advertising a "natural language" search box that automatically inserts Boolean operators between entered words; these systems often feature natural

language interfaces with partial matching retrieval techniques. Dialog's TARGET, LEXIS/NEXIS's FreeStyle, and Westlaw's WIN are a few that seem to have received the most of the media's attention, and therefore, have been the most tested in recent studies. The results of these studies have suggested that certain types of queries, namely broader questions of "about-ness," retrieve more relevant results with natural language searching than with traditional Boolean searching techniques, whereas queries relating to specific pieces of information may be better served by Boolean searches (Feldman, 1996; Paris & Tibbo, 1998; Pritchard-Schoch, 1995; Tenopir & Cahn, 1994; Tomaiuolo & Packer, 1998).

The focus of these studies has chiefly been on commercial databases of full text articles or abstracts on law or medicine. The implications of natural language searching in full text databases that serve academic users, undergraduate students in particular, have largely been ignored by researchers. This proposed study seeks to investigate whether natural language searching may turn up more relevant articles for undergraduates, especially if the search question is general in nature, than Boolean searching does.

Studies on undergraduate students' information seeking behaviors show that the majority of their searches are for classes in which their professors ask for a research paper covering a broad topic (Leckie, 1996); students then tentatively approach databases offered by the library to search for journal articles on their topic, only to find that Boolean searching is the only search option available in many of the databases to which the library subscribes. Natural language searching, with the added advantage of being

easier to navigate for the novice user, may be more helpful to students with general information needs.

This thesis presents a field experimental study that uses the WilsonWeb databases, available through San José State University, to compare Boolean searching against natural language searching in the context of undergraduates' academic research. A literature review will examine the current status and development of NLP in information retrieval as well as natural language interfaces; similar studies testing the relative advantages of Boolean over natural language searching will also be reviewed to provide context for this study.

Literature Review

NLP in Information Retrieval

Interest in NLP in information retrieval began in the 1960s when researchers realized that computational linguistics techniques could be applied to the problem of document indexing and retrieval; systems would be able to “compute the appropriate relationships between the parts and subparts of text in textual databases and queries ... to identify the meaningful units of these texts and to determine how they relate to one another” (Doszkocs & Weinberg, 1988, p. 127). Unfortunately, progress in this area was halted in the 1970s and 1980s when funding was denied because the technology necessary to support such systems was thought to be “beyond then available computational capabilities” (Liddy, 1998, p. 14). Interest resurged in the early 1990s when advances in technology made research in this area more feasible. Today, researchers see an optimal, pure NLP system as one that should be able to use linguistic analysis to parse a user query and analyze it as a reference librarian would do with a patron’s question, and then retrieve relevant documents using the same level of linguistic analysis (Liddy, 1997).

Several levels of linguistic processing have been studied by researchers in IR.

Liddy (1998) categorizes them as follows:

- Morphological – componential analysis of words, including prefixes, suffixes, and roots;

- Lexical – word level analysis including lexical meaning and part of speech analysis;
- Syntactic – analysis of words in a sentence to uncover its grammatical structure;
- Semantic – determination of possible meanings of a sentence, including disambiguation of words in context;
- Discourse – interpretation of structure and meaning conveyed by texts larger than a sentence;
- Pragmatic – understanding the purposeful use of language in situations, particularly those aspects of language which require world knowledge (Liddy, 1998, p. 14).

Each level in the hierarchy builds upon the previous levels. For example, understanding a text at the syntactic level would require analysis at the lexical and morphological levels, while analyzing text at the semantic level would require understanding at the syntactic, lexical, and morphological levels, and so on. A system would have to be able to analyze both a query and the document text at all linguistic levels for it to be considered a pure NLP system. In the past, the focus of research in this area had been on the lower linguistic levels, and NLP techniques resulting from such research include truncation (processing at the morphological level), stop word lists (processing at the lexical level), and extraction of noun phrases (syntactic level) (Warner, 1990). Only recently have system developers produced any kind of applied processing at the semantic or pragmatic levels, and such processing tends to be very limited and subject

specific. Needless to say, such features are unavailable in systems used by the average university user (Liddy, 1998).

Natural language interfaces, on the other hand, have enjoyed a much more successful run. This is partly due to the fact that designing a user-friendly NL interface does not involve complex algorithms for linguistic parsing like that necessary in NLP. In fact, Doszkocs and Weinberg (1988) reveal that many of the original NL interfaces allowed the user to enter sentences or phrases in free form for searching, but the actual processing was done by automatic insertion of Boolean ANDs or ORs between entered words, not by NLP. With recent advancements in technology, many systems now are able to not only present an NL interface to the user but also apply basic linguistic analyses in the form of automatic stemming, phrase identification, and checking for stop words (Tenopir, 1994).

In addition to providing a natural language interface, many systems are now combining various levels of NLP with other techniques in an effort to improve performance in information retrieval. Pritchard-Schoch (1993) grouped IR models into three classes:

- Exact match model– Boolean logic is essentially an exact match search; specific terms are identified and searched within the database, and only documents containing those terms are retrieved

- Vector space model – this model uses word frequencies to determine term weights in retrieved documents and computes query-document similarities with the cosine measure for the purpose of relevance ranking;
- Probabilistic model – this model ranks retrieved documents in order of their predicted probabilities of relevancy to a given query.

An example of a database that utilizes NLP and vector space and probabilistic models in retrieval (and feature an NL interface) is Westlaw's WIN (West Is Natural). Westlaw processes queries and retrieves documents using four processes, as explained by Pritchard-Schoch (1993) below:

- 1) Identification of key concepts, i.e., words and phrases;
- 2) Removal of stop words such as "to," "the," etc.;
- 3) Application of linguistic techniques to determine roots and/or expand on roots;
- 4) Performance of a statistical comparison of retrieval (Pritchard-Schoch, 1993, p. 35).

The majority of commercial systems offered today are similar to Westlaw's WIN in that they do not apply all levels of NLP in query parsing or text retrieval; while the actual search algorithms utilized by commercial systems are proprietary, we do know that at the most basic level many of these algorithms employ partial matching retrieval techniques that use weighted terms and word frequencies to calculate a score for relevancy ranking (Paris & Tibbo, 1998; Quint, 1994). From here on, the phrase "natural language searching" will be used to describe searching in any system that implements a

combination of natural language interface, query parsing with some degree of linguistic analysis on at least the syntactic level, and usage of partial matching techniques like vector space and/or probabilistic models in document retrieval.

Current Research and Findings

Pritchard-Schoch (1993) stated that:

in general, the exact match models work well for known-item searching and for bibliographic and field searching. The vector space and probabilistic models have shown better performance levels for more general searching of full-text files, or for the untrained searcher. (p. 34)

There have been several studies that compared the results of natural language searches in databases that use NLP with vector space and/or probabilistic models against

Boolean/exact matching searches. Feldman (1996) reviewed DIALOG, TARGET, and DR-LINK; Tenopir and Cahn (1994) compared DIALOG and FreeStyle; and Tomaiuolo and Packer (1998) took a brief look at EBSCOhost's Academic Search Full-Text Elite.

While the aforementioned three studies were cursive in analysis, Paris and Tibbo (1998) fully documented an investigation into partial versus exact matching retrieval systems through a comparison between Boolean and natural language searching in the FreeStyle system.

Paris and Tibbo's (1998) study was unique in that they selected a small database in which all the documents were known and possible to be judged for relevancy, thus enabling them to calculate both precision and recall scores for the system. In contrast, the other three studies queried databases in which the indexed documents were not known and therefore recall scores were not determinable. In addition, due to the

proprietary nature of the algorithms used by IR systems, all studies were black-box evaluations (in contrast to glass-box evaluations) which, according to King (1996), means that “access to intermediate results is denied, so that the evaluator can work only with the outputs produced by particular inputs and perhaps a tertium comparationis in terms of a predefined expected output” (p. 77). Analyses into the level of linguistic processing performed by the systems in query parsing and retrieval were purely speculations.

Each of these studies also had various limitations and further research is needed to reveal possible advantages of using natural language searching in the academic environment. As previously noted, these studies, with the exception of Tomaiuolo and Packer (1998), were conducted on databases that were largely subject specific and not available to the average university student for research. Specifically, the study by Paris and Tibbo (1998) assessed an incredibly narrow subject database – the Cystic Fibrosis database, available through MEDLINE, containing a total of 1,679 articles from the period of 1974-1979. Though Tomaiuolo and Packer studied EBSCOhost’s Academic Search Full-Text Elite, a database that is widely available to colleges and universities, the informality of the study deems their findings less credible, or at best tentative and exploratory in nature.

Studies by Feldman (1996) and Tenopir and Cahn (1994) suffered from similar weaknesses. While there was no limit to the number of retrieved documents to be judged for relevancy, the number of queries that were run through the systems was too small to produce generalizable results. In a study of DIALOG, TARGET, and DR-LINK, Feldman (1996) tested a total of four questions that were transformed into Boolean,

keyword without Boolean operators (called unadorned relevance by Feldman), and natural language searches. Tenopir and Cahn's (1994) test of TARGET and FreeStyle only examined the results of six questions, though there was a sense of realism as they were all real questions gathered from reference librarians. The limited number of questions was an admitted disadvantage in their study; the authors wrote that "an in-depth comparison of these Boolean search engines with relevance search [ranking] techniques requires testing real questions and searches" and that "[t]his should be done over time by many searchers – we have just scratched the surface" (Tenopir & Cahn, 1994, p. 40).

Despite these drawbacks, several interesting conclusions have emerged from this body of research:

- Searching that utilizes partial matching techniques and lower level NLP may be better for vague questions or searches on broader subjects (Feldman, 1996; Tenopir & Cahn, 1994);
- The search results may be better when the query involves two or more "concepts of unequal weight" (Tenopir & Cahn, 1994);
- Boolean search is not superior to NLP-based or partial matching retrieval, but rather, "different queries demand different retrieval mechanisms" (Paris & Tibbo, 1998, p. 189);
- Because there is often little overlap between retrieved documents, both search techniques should be employed to find the maximum number of relevant

documents for the user (Feldman, 1996; Paris & Tibbo, 1998; Tenopir & Cahn, 1994; Tomaiuolo & Packer, 1998).

This study attempts to bridge the gap in current research on the potential of natural language searching in full text databases in the context of undergraduate students' research in an academic library.

Definition of the Research Problem

Although the implementation of a full NLP system has not yet been possible for the majority of commercial systems and databases, the availability of systems that offer natural language searching using a combination of natural language interfaces, query parsing at the syntactic level using lesser degrees of NLP, and partial matching techniques for document retrieval may still be a valuable tool for database searchers. The main purpose of this study is to investigate whether natural language searching is more effective than Boolean searching when a student has a broad research question, with the term “broad” defined here as queries that include two or fewer concepts. For example, the query “What is the effect of PCBs on fish?” has two concepts, while the query “Should the state provide emergency medical services for illegal immigrants?” has three (Tenopir & Cahn, 1994, p. 36). Secondary purposes include whether natural language searching should be considered as a desirable option in addition to Boolean searching in databases that serve academic users.

Major Research Questions

- Would a student’s broad query retrieve more relevant documents with a natural language search than with a Boolean search?
- Do more specific queries (or queries with three or more concepts) retrieve more relevant results with Boolean searches?

- Do natural language searches retrieve results that are unique compared to those from Boolean searches (i.e., how much overlap is there between retrieved documents)?
- Does performing both a natural language search and a Boolean search increase the total number of unique relevant documents retrieved?

Minor Research Questions

- Does the addition of a natural language search option assist students in gathering information for their research needs, or does it merely add an extra and ultimately unnecessary step in the research process?
- Should natural language searching be suggested to undergraduate students as an alternative or complementary to Boolean searching?

Statement of Major Research Hypothesis

- H1 Broad queries (queries containing two or fewer concepts) retrieve a greater number of relevant results with natural language searches than with Boolean searches.
- H2 Using both natural language search and Boolean search increases the total number of relevant results.

Methodology

Research Design

This study is set up as a simple field experiment in essence, with NLP/Boolean searching on corresponding database systems as the treatment and the same set of real-life questions (statements of information needs) collected from undergraduate students as experimental subjects.

Pilot Study of Relevance Judgment Guide

A pilot study was conducted during the Fall 2005 semester in order to identify potential problems or unanticipated consequences in the execution of the study. Particular attention was paid to the validity of the instrument chosen to measure document relevancy, and to the number of retrieved results students could accurately and comfortably review.

The pilot study consisted of a group of twenty students recruited from the library as they were entering or exiting. Execution of the pilot study was the same as the actual study, except for the pilot study that in addition to judging citations as “relevant” or “not relevant,” students were also asked to circle each applicable relevancy criteria.

The four chosen criteria presented to students for consideration in making relevance judgments were taken from Maglaughlin and Sonnenwald’s (2002) work. In their study, they found that there was a pattern in criteria used by students to judge relevance. Four criteria, namely, breadth, currency, author (credibility), and subject matter (topic appropriateness), were identified in ten different studies on relevance judgments. In the pilot study, the researcher suggested to the participating students that

these four criteria be considered in relevance evaluation. The majority of the pilot study students (sixteen out of twenty) consistently circled subject matter, currency, and breadth as criteria that effected their relevance judgments. In consideration of the positive responses from the pilot study students, participants in the actual study were asked to use the same four criteria in making relevance decisions.

Results from the pilot study led to two minor changes to the initial proposed procedure. Searches were originally restricted to only full text articles in order to ensure that students could review abstracts and articles to make the best relevance judgments possible. However, it was discovered that limiting the results to only full text articles meant that the search would miss out many articles of which full texts could be found in other databases through SFX's Get Text linking technology. In order to optimize search results for the students, the full text restriction was lifted in the actual experiment. In addition, the option of "possibly relevant" was added to the choices of "relevant" and "not relevant" for those citations that did not include an abstract or full text article and could not produce an accurate relevance rating. Most students, however, were able to make judgments based on the information given by the database system, and as a result the "possibly relevant" option was rarely used.

Operational Definitions

Variable: Query properties

1. Broad query: Query containing one or two concepts; i.e., "What is the effect of PCBs on fish?" (Tenopir & Cahn, 1994, p. 36).

2. Narrow query: Query containing three or more concepts; i.e., “Should the state provide emergency medical services for illegal immigrants?” (Tenopir & Cahn, 1994, p. 36).

Variable: Document relevancy

1. Relevant document: End-user defined. A relevant document will be labeled “Yes” on the Relevance Rating Form by the user.
2. Non-relevant document: End-user defined. A non-relevant document will be labeled “No” on the Relevance Rating Form by the user.
3. Possibly-relevant document: End-user defined. A possibly-relevant document will be left unmarked on the Relevance Rating Form.

Variable: Number of documents retrieved

1. Total relevant documents: Total number of documents retrieved that are judged “relevant” by the end-user.
2. Total retrieved documents: Total number of documents retrieved by the system for a given search.
3. Total unique relevant documents: Total number of documents that were judged “relevant” by the end-user that are retrieved using one type of search but not the other.

Statistical Hypotheses

H₀: There is no significant difference in the number of relevant documents retrieved for broad queries between natural language searches and Boolean searches.

$$\mu_N = \mu_B$$

H_1 : The number of relevant documents retrieved for broad queries using natural language searches will be greater than number of relevant documents retrieved for broad queries using Boolean searches.

$$\mu_N > \mu_B$$

Where μ_N is the average number of relevant documents retrieved with natural language searches and μ_B is the average number of relevant documents retrieved with Boolean searches.

Research Population and Sampling

This experimental study was conducted in the Spring semester of 2006 at San José State University, a large public university with an undergraduate student population of over 20,000 students, and a graduate student population of 6,100 students. The university is divided into seven colleges (followed by the number of undergraduate students in each college): Applied Sciences and Arts (4,272), Business (4,968), Education (1,510), Engineering (2,749), Humanities and the Arts (3,405), Science (1,841), Social Sciences (2,572), and Undeclared/Unclassified (1,181).

The research sample was queries drawn from San José State University undergraduates who were conducting research for a paper, project, or assignment that required library research. Random sampling of real-life queries is required for this experiment. However, because students were needed to make relevance judgments for

their own queries, students (and thus their queries) were randomly selected instead, and participation was voluntary.

Students were recruited outside the university library entrance/exit as well as through their classes. Although the original plan was to obtain an electronic mailing list from each of SJSU's seven colleges, this proved to be impossible because of privacy concerns with the colleges. Instead, an e-mail explaining the study and requesting for assistance in recruiting participants was sent to each of the university subject librarians, who then forwarded the e-mail to their respective department faculty. Students were thus given contact information for the study by their professors in class.

Selection of Database

WilsonWeb was chosen as the test database for this study for three reasons: (1) it is readily available to all currently enrolled San José State University students; (2) it offers NL searching in addition to Boolean searching; and (3) it contains full-text databases that cover the majority of subjects taught at SJSU.

WilsonWeb utilizes the Verity search engine algorithm in its natural language searches; Verity parses entered queries at the syntactic level (utilizing morphological/syntactic level NLP) and calculates term weights and relevancy rankings for retrieved documents (Oldenkamp, 2003; Quint, 2002).

The following subject databases are offered through WilsonWeb: Art Full Text, Education Full Text, Library Literature and Information Science Full Text, Social Sciences Full Text, Biological and Agricultural Index Plus, General Science Full Text, Business Full Text, Humanities Full Text, Readers' Guide Full Text, and OmniFile Full

Text (essentially a compilation of all of the above full-text subject databases). In an effort to prevent students from being overwhelmed by the number of retrieved documents that they have to evaluate for relevancy, only the top 50 retrieved documents were presented for judgment. Each student was asked to select the most appropriate database for every individual query – for example, a question about the influence of John Locke’s philosophy on modern political theory would only be searched in Social Sciences Full Text.

Procedure

Participating undergraduate students were recruited from the university library as well as in their classes during the Spring 2006 semester. Recruitment from the library was done by obtaining permission to set up a table outside the entrance/exit of the library with information about the study and contact information. Recruitment in classes was done via e-mail to professors, who were asked to give contact information to their students. Virtually all the students that responded to the request for participants volunteered to submit queries for the study. Interested students were asked if they were doing research for a paper or a project for class, and were also asked for their research topic in order to ensure that the subject was covered and therefore searchable in any of WilsonWeb’s databases. Students were selected for the study if both criteria were met and the Human Subjects Consent Form was signed.

Each identified student was asked to write down his/her research question that he/she was working on in two forms, one as plain English statement and the other in Boolean expression. The researcher took the queries and conducted the searches on

WilsonWeb one by one; in other words, each question was run through the subject-appropriate WilsonWeb database twice, once as a natural language search and again as a Boolean search. The total number of queries tested was 96.

The top 50 retrieved citations were printed out for each query. WilsonWeb allows the user to specify which fields should be displayed in the retrieved citation list, so each citation included the following information: article title, author, peer-reviewed journal, journal name, source, publication year, abstract, subject headings, database, accession number, and the persistent URL for the article. Article title, author, journal name, source, publication year, and abstract were selected to help students make their relevance judgments. Although not directly pertinent to relevance judgments, information about peer-reviewed journals, subject headings, accession numbers, and persistent URLs were included in the printouts in order to assist the students if they wanted to expand or continue their research in other databases or catalogues.

Relevance Judgment

The printouts of search results were returned to the students for evaluation. Students were then asked to judge each retrieved citation as “relevant,” “not relevant,” or “possibly relevant,” and complete a form (see Appendix A) as a way of recording their judgments. Often in cases where real life queries are used to test systems, researchers collect the queries from users, but the relevance of returned results are judged by the researchers rather than the user themselves (Tenopir & Cahn, 1994; Tomaiuolo & Packer, 1998). The practice of using researchers to make the relevance judgments has been questioned. In their experimental setup for retrieval system comparisons, Paris and Tibbo

(1998) argue for the importance of realism when finding queries to test and evaluating results, leading them to not only use real life, self-generated queries from scientists and physicians (testing a medical research database) in their study, but also to use the same group of scientists to make the relevance judgments for the retrieved results sets. This study of academic users strives for a similar level of realism by having the participating students make their own relevance judgments.

Whether students' search questions generated from research topics assigned by their professors or from course assignments (as in the case of this study) constitute "imposed queries" is a question that needs some clarification. In examining the differences between reference services to users with self-generated questions and those with imposed questions, such as students with class assignments, Gross (1999) warns of the probabilistic nature of relevance judgments made by the latter group. Pertinent to this study is Gross' assertion that students with writing assignments are essentially secondary agents that are searching for information for the imposer, namely the professor, and thus are not likely to be able to make accurate relevance judgments. However, for the purposes of this study, judgments from the students who will actually be using the articles in completing their assignments would still be more accurate than the tertiary relevance judgments that would have been made by the researcher, although not directly comparable to Paris and Tibbo's (1998) work. In addition, one of the benefits of allowing the students to browse retrieved citations, relevant or not, is a better sense of the topic at hand and an indication of the need for the searcher to narrow or broaden a search accordingly. Additional research help was offered to students as a benefit to participating

in the study, and students responded positively to using the retrieved citations to find keywords, concepts, and subject headings that they otherwise would not necessarily have searched.

Although the relevancy of retrieved documents is ultimately determined by the end-user and therefore subjective, students were asked to consider the following criteria when making their decisions: topical appropriateness, currency, breadth, and creditability (see the section *Pilot Study* for discussion on these criteria).

The resulting queries along with the respective lists of retrieved document citations and relevance judgments were collected and then labeled with the number of concepts in the query. The subject database used for each query was indicated on the top of every list. All forms and data collected from participants were entered into SPSS as numerical data for quantitative analysis.

Sample Queries

A total of 48 queries and responses were collected, of which 46 were usable. The majority (35) of these usable queries were from the Social Sciences, while 4 fell under the subject category of Art, 4 under Humanities, 1 under Education, 1 under General Science, and 1 under Business. The number of concepts, however, was relatively evenly distributed. Twenty-one queries have 2 or fewer concepts and thus are deemed as “broad queries,” specifically, 5 with one concept, 16 with two concepts. Twenty-five queries have 3 or more concepts and therefore are treated as “narrow queries,” specifically, 23 with three concepts, 1 with four concepts, and 1 with five concepts. Examples of broad queries include searches on elder abuse, irrigation in Mesopotamia, and access to leisure

activities by the disabled. Narrow queries included searches on the history of gay community groups, population control in India and China, and the Filipino immigration experience between 1920 and 1965 in Hawaii and California (see Table 1).

Table 1

Query Examples

Query Type	NL Search	Boolean Search
Broad queries (one to two concepts)	Elder abuse	elder AND abuse
	Development of irrigation in Mesopotamia	irrigation AND Mesopotamia
	Access to leisure activities by the disabled	leisure AND disabled
	Themes in J. D. Salinger's writing	themes AND Salinger
Narrow queries (three or more concepts)	Population control practices in India and China	(India OR China) AND population AND control
	Filipino immigration experience between 1920 and 1965 in California and Hawaii	Filipino AND immigration AND California AND Hawaii
	What are the environmental impacts of oil drilling in Alaska's Arctic National Wildlife Refuge	environment AND oil AND Alaska drilling AND Arctic national wildlife refuge

Results

Broad Queries

Paired-samples t-tests were performed to measure whether there was a statistical difference in the number of relevant documents, total retrieved documents, and unique relevant documents retrieved using a natural language search versus a Boolean search in broad queries. According to the convention of empirical research, statistical significance level was set at .05. The number of total relevant documents, total retrieved documents, and total uniquely relevant documents (relevant documents that were unique to either a Boolean or NL search) were also analyzed for both types of searches to determine the average number of articles found for each search.

Queries were divided into two groups for analysis in SPSS by using the select cases function; one group consisted of broad queries while the other group consisted of narrow queries. The first group to be analyzed was the broad query group, with a total of 21 out of the 46 collected queries containing two or fewer concepts ($n=21$). Within this group, the mean was calculated for three groups of data: total number of relevant documents, total number of retrieved documents, and total number of unique relevant documents (see Table 2). The first group, total number of relevant retrieved articles, had a mean of 7.71 for NL searches and a slightly smaller mean of 6.76 for Boolean searches. For the second group, natural language searching produced a significantly larger number of total retrieved articles than Boolean searching, with a mean of 46.43 to Boolean's mean of 20.81. The third group, total number of unique relevant articles, was similar to

the other two groups in that NL searches had a greater mean of 4.62 articles compared to the mean of 3.43 for Boolean searches.

Table 2

Means for Total Relevant Articles, Total Retrieved Articles, and Total Unique Relevant Articles: Broad Queries

Group	Mean	Std. Deviation
1 – Total number of relevant retrieved articles		
Natural Language	7.71	7.682
Boolean	6.76	6.066
2 – Total number of retrieved articles		
Natural Language	46.43	9.490
Boolean	20.81	21.621
3 – Total number of unique relevant articles		
Natural Language	4.62	6.054
Boolean	3.43	4.094

Note. (n=21)

The paired-samples t-tests for this data showed that there was no significant difference in either total number of retrieved relevant items or total of unique relevant items ($p = 0.547$ and $p = 0.448$ respectively). Statistically significant difference was found on in the total number of retrieved articles between NL and Boolean searching ($t=5.682$, $p<.05$), as shown in Table 3.

Table 3

Results, Paired Samples t-Test: Broad Queries

Group	t	df	Sig. (2-tailed)
1 – Total number of relevant retrieved articles (NL-Boolean)	.613	20	.547
2 – Total number of retrieved articles (NL-Boolean)	5.682	20	.000
3 – Total number of unique relevant articles (NL-Boolean)	.774	20	.448

Narrow Queries

A paired-samples t-test was also performed to measure whether there was a significant difference in the number of relevant documents retrieved using a natural language search versus a Boolean search in narrow queries. Statistical significance level was set at .05. The number of total relevant documents, total retrieved documents, and total unique relevant documents (relevant documents that were unique to either a Boolean or NL search) were also analyzed for both types of searches in order to determine the average number of articles found for each search.

As with the broad queries, the relevance responses for the narrow queries were divided into three different groups – (1) total number of relevant retrieved articles, (2) total number of retrieved articles, and (3) total number of unique relevant articles. Table 4 summarizes the means for the three groups.

Table 4

Means for Total Relevant Articles, Total Retrieved Articles, and Total Unique Relevant Articles: Narrow Queries

Group	Mean	Std. Deviation
1 – Total number of relevant retrieved articles		
Natural Language	5.48	7.693
Boolean	10.44	9.368
2 – Total number of retrieved articles		
Natural Language	45.44	12.534
Boolean	22.88	18.406
3 – Total number of unique relevant articles		
Natural Language	3.28	4.614
Boolean	8.24	8.599

Note. ($n=25$).

Paired t-tests for these three groups of narrow queries showed that there is a significant difference between NL and Boolean searching in the total number of relevant retrieved items ($t = -2.294, p = 0.02$), the total number of retrieved articles ($t = -2.294, p = 0.02$), and the total number of unique relevant items ($t = 4.695, p < 0.001$).

Table 5

Results, Paired Samples t-Test: Narrow Queries

Group	t	df	Sig. (2-tailed)
1 – Total number of relevant retrieved articles (NL-Boolean)	-2.294	24	.020
2 – Total number of retrieved articles (NL-Boolean)	4.695	24	.000
3 – Total number of unique relevant articles (NL-Boolean)	-2.294	24	.020

Effects of Subject on Retrieved Articles

An attempt was made to investigate possible interaction of subject or topics with search types (NL vs. Boolean) and query types (narrow vs. broad) in the number of relevant articles, retrieved articles, and unique articles using a multivariate analysis of variance (MANOVA). In order to run such an analysis, three new columns were added to the original SPSS data sheet to encode differences between NL and Boolean search types in the total number of relevant documents retrieved, the total number of retrieved documents, and the total number of unique relevant documents retrieved. However, the small sample size prevented adequate distribution of sample queries into experimental cells. Specifically, 35 of the queries collected were categorized as Social Science (76%), while 4 were Art (9%), 4 were Humanities (9%), and 1 each for Education (2%), General Science (2%), and Business (2%). Consequently, such an analysis of interaction effects was meaningless.

While the imbalance of subjects could be due to the fact that certain classes – specifically, Social Science, Art, and Humanities – may require the most writing assignments and thus would have been more likely to respond to the call for participants, it is undeniable that all the areas of study offered by San José State University require library research at some point in the curriculum. It is therefore recommended that future studies make a conscious attempt to actively gather queries from underrepresented groups.

Discussion

The results of this field experimental study produced no supporting evidence for the original hypothesis that NL systems would retrieve a greater number of relevant results with broad query search than narrow query search. Though natural language searching retrieved significantly more articles than Boolean searching for broad queries, it failed to produce either significantly more relevant articles or significantly more unique relevant articles than Boolean searching. In essence, with broad queries, a Boolean search would yield the same results as an NL search.

With narrow queries, however, Boolean searching did significantly better than NL searching. The total number of relevant articles retrieved and the total number of unique relevant articles retrieved were much greater with Boolean searching. However, as with searches with broad queries, NL searching retrieved a greater number of total articles than Boolean searching.

Implications for Student Searching

At first look, these research findings seem to suggest that a student with a broad query may choose to conduct a basic NL search and be confident that the results would be comparable to the results from a more complicated Boolean search. However, a closer look at the statistics reveals that the total number of articles retrieved is considerably larger with an NL search ($M=46.43$) than with a Boolean search ($M=20.81$); a student conducting a search with a broad query using natural language processing would retrieve an average of 7.71 relevant citations out of a total of 46.43 citations, at an average precision rate of 16%. In comparison, a student conducting a broad query search with

Boolean searching would only need to review an average of 20.81 citations to find 6.76 relevant citations, at an average precision rate of 32%. In other words, students would have to review more than double the number of articles on average to find one more relevant article if the search is conducted with an NL query.

It is also important to note that in order to prevent the students from being overwhelmed by the size of the citation lists, students in this study were asked to review only the first 50 retrieved citations for each search. This cutoff is reflected in the statistical analysis, where the number of total retrieved articles never exceeded 50 and the average number of total articles retrieved was 46.43. In reality, natural language searches consistently retrieved far more than 50 citations, and sometimes even thousands of citations. Though it is unlikely in reality that students would review thousands of citations, the fact that many of the NL searches were able to retrieve so many articles means that the precision rate must in actuality be lower than the 16% observed in this study. Natural language searching's tendency to retrieve a greater number of less relevant citations was well documented by Paris and Tibbo (1998) and by Tamaiuolo and Packer (1998), who noted that users should be willing to review a greater number of unrelated items when conducting an NL search.

Taking into consideration the amount of time and effort required to review retrieved articles, it would be arguably more time efficient for a student to perform a Boolean search rather than an NL search. The claims by Feldman (1996) and Tenopir & Cahn (1994) that NL searching is better for vague questions or broad subject searches are not supported by the findings of this study.

The implication of these experimental results with narrow queries is much more straightforward. Students with narrower queries of three or more concepts would be much more successful in finding relevant articles with Boolean searching, which retrieves about double the number of relevant articles that NL searching did: 10.44 relevant articles in Boolean searches in comparison to 5.48 relevant articles in natural language searches. In addition, relevant articles found with Boolean searching were more likely to be found only in Boolean searches and not in NL searches. While it is clear that Boolean searching with narrow queries is more successful, we have to look to the users themselves to answer the question of whether it is advisable for librarians to suggest that students use a natural language search in addition to a Boolean search in order to find the greatest number of relevant articles.

Feldman (1996), Paris and Tibbo (1998), Tenopir and Cahn (1994), and Tamaiuolo and Packer (1998) all advise using both search techniques for the purpose of finding the greatest number of relevant articles, and indeed running both types of searches would technically increase the total number of relevant articles found. Whether it is *worthwhile* to run both types of searches would depend on the amount of effort the student is willing to spend. Once again, we have to look at the number of relevant articles retrieved in comparison to the total number of articles retrieved. The ratio of unique articles to the total number of articles retrieved for a narrow query with a natural language search is 3.28/45.44, meaning that conducting an NL search in addition to a Boolean search would return an average of three more relevant articles that were not found with the Boolean search but requires the student to review 45 more retrieved

citations. In addition, as mentioned previously, in this study only the first 50 retrieved citations were printed out for relevance judgment. In reality, many of the NL searches for narrow queries retrieved more than 50 citations, which would increase the total number of citations retrieved on average. It seems hardly worth the time and effort to review all the citations retrieved from an NL search when a more effective method to increase the recall may be to use another search technique. One common way of finding additional literature is to use the citations that are referenced in relevant books and articles, a technique made simple by databases like ISI's Web of Science. In addition, students could use search strategies commonly taught in library instruction, such as employing subject headings provided by the database or using alternate keywords and concepts found in relevant articles to improve their relevant retrieval rate.

It seems that natural language searching as a search strategy would only be preferable to Boolean searching if the user has a broad query and needs to retrieve a large number of citations and abstracts in order to narrow his or her topic. This type of exploratory searching was briefly discussed by Feldman (1996), who stated that from a user's perspective:

being able to browse around my stated query brought me much more knowledge than having answers confined to what I had asked ... perhaps this is the crux of the issue: people who are searching for information often can't ask for what they need. (Feldman, 1996, p. 73)

In this study, natural language searches yielded a high number of total retrieved citations, which certainly suggests that users are getting what they searched for and more through

NL searching. The user, however, must be open to spending the time and cognitive effort to review additional citations.

Implications for Database Designers and Librarians

The level of processing currently employed by WilsonWeb in its natural language search option is not an extremely effective way for undergraduate students to search, and the system's effectiveness decreases as the users' queries become narrower and more focused. What does this mean for database designers and librarians? While the lack of the NL search option in most of San José State University's online databases indicates that NL has fallen out of favor since the 1990s, the existence of the single search box in most Internet search engines suggests that many people expect to retrieve relevant results based on a single NL query. Today's undergraduates come into the university library with considerable experience with Internet searching but possessing few skills for scholarly database searching – effective keyword searches using subject headings and field limiters, for example. Online databases and their subscribing libraries can help students transition and learn to develop searching skills by supporting library instruction.

The findings of this study support, if nothing else, the need for students to learn and develop search strategies and critically analyze content in order to retrieve the most relevant and appropriate articles for their information need. These skills contribute to searching as an iterative process, one that requires continuous fine tuning of queries as information needs are met and changed. This study only looked at the first step of the research process, where the user enters the initial query into a database. The searching process presumably continues even after the initial search is completed, as many of the

students that participated in the study were interested in using the retrieved citations lists to continue or expand their search in WilsonWeb and also in other databases, regardless of how many relevant citations were found in the initial search.

Ultimately, the future of database searching lies in the progress to be made in the fields of human/computer interaction and information storage and retrieval technology. Emerging technologies such as Metalib and SFX, along with other cross-database retrieval systems, will likely make database searching easier for the average user by offering one interface laid over several other database interfaces, so that the user only has to search one interface to access many databases at once. Although it is unclear whether meta-searching can support natural language processing, meta-searching interface is unlikely to increase the success of a search without the user knowing about the organization of information in the systems, which in turn requires instruction. Advancements in human/computer interaction would mean user-friendly interfaces for cross-database technologies and individual databases that could offer both Boolean or NL processing. An easier interface does not necessarily mean that finding relevant information will be less time consuming or cognitively taxing, as evidenced by the number of retrieved articles found in an NL search compared to the number found in a Boolean search. Effective library instruction would have a huge impact on students' maximum exploitation of the potential of any searching system.

Suggestions for Further Study

One of the main drawbacks of this study was the small sample size. Although the homogeneity of an undergraduate population makes the requirement of a very large

sample size less essential to producing accurate and statistically significant research findings, a larger sample representative of the diversity of the school with regards to majors/subject areas would have allowed analysis to greater depth of the interaction of subjects and concepts on the number of retrieved, relevant, and unique articles. Further research with larger sample sizes and coverage of more subjects may show that one type of searching may be more successful than another for certain subjects.

Although NL searching was found not to be as effective as Boolean searching in this particular study, the applicability of the findings is limited because the results are based on the searching algorithm of one database. Oldenkamp (2003) and Quint (2002) wrote that Verity, the program WilsonWeb uses for NL processing in their databases, parses entered queries at the syntactic level (utilizing morphological/syntactic level NLP) and calculates weighted terms and relevancy rankings for retrieved documents. It would be interesting to conduct further research with systems that employ more sophisticated linguistic processing to see how well these systems could meet the research needs of undergraduates.

Conclusion

This field experiment found no statistically significant difference between a natural language search and a Boolean search in the context of undergraduate students' researching on a broad topic. If the student is willing to put in extra effort to review a longer list of retrieved citations, then a natural language search may be preferable to a Boolean search, only because it does not require the mastery of keyword searching skills. When researching a narrow topic, however, Boolean searching should be recommended.

Even though natural language processing has not reached the level of effectiveness and sophistication as many had hoped, today's natural language searching systems can still be "another weapon in the good searcher's arsenal" (Tenopir & Cahn, 1994, p. 46). A more interesting observation from this study is that students seem completely unaware of the fact that there exist other kinds of search options in addition to keyword searching. When the purpose of this study was explained to potential participants, many of them expressed surprise that the default search screen, which is usually set to keyword searching, could be changed to a different type of searching function. Hopefully, this study will increase users' awareness of alternative searching techniques that may have been overlooked in the past due to the prevalence of Boolean searching. Optimally, this awareness would be kindled in three groups of users: one group being students themselves who may learn another road to searching for what they need, another group being reference librarians, and the third group being researchers and developers of information retrieval systems who seem to have lost some interest in NLP in full text databases in recent years, as indicated by the decreasing amount of literature

published on NLP in database systems since the late 1990s, although literature on NLP in Web search engines has increased (Feldman, 2000).

References

- Arnold, S., & Rosen, L. (1993). Bye bye, Boolean: Natural language and electronic information retrieval. *Searcher*, 1(5), 30-38.
- Doszkocs, T. E., & Weinberg, B. H. (1988). Natural language interfaces for information retrieval. In J. A. Benson & B. H. Weinberg (Eds.), *Gateway software and natural language interfaces: Options for online searching* (pp. 123-134). Ann Arbor, MI: Pierian Press.
- Feldman, S. (1996). Testing natural language: Comparing Dialog, Target, and DR-Link. *Online*, 20(6), 71-74.
- Feldman, S. (2000). Find what I mean, not what I say. *Online*, 24(3), 49-56.
- Gross, M. (1999). Imposed versus self-generated questions. *Reference & User Services Quarterly*, 39(1), 53-61.
- King, M. (1996). Evaluating natural language processing systems. *Communications of the ACM*, 39(1), 73-80.
- Leckie, G. J. (1996). Desperately seeking citations: Uncovering faculty assumptions about the undergraduate research process. *Journal of Academic Librarianship*, 22(3), 201-208.
- Liddy, E. D. (1997). Natural language processing for information retrieval and knowledge discovery. In P. A. Cochrane & E. H. Johnson (Eds.), *Visualizing subject access for 21st century information resources* (pp. 137-147). Urbana-Champaign, IL: University of Illinois.
- Liddy, E. D. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science*, 24(4), 14-16.
- Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, 53(5), 327-342.
- Oldenkamp, D. M. (2003). The new Wilsonweb. *Searcher*, 11(3), 17-25.

- Paris, L. H., & Tibbo, H. R. (1998). Freestyle vs. Boolean: A comparison of partial and exact match retrieval systems. *Information Processing & Management*, 34(2/3), 175-190.
- Pritchard-Schoch, T. (1993). Natural language comes of age. *Online*, 17(3), 33-43.
- Pritchard-Schoch, T. (1995). Comparing natural language retrieval: Win & Freestyle. *Online*, 19(4), 83-87.
- Quint, B. E. (1994). Connect time: The artifices of natural language searching. *Wilson Library Bulletin*, 69, 60-61.
- Quint, B. E. (2002). H. W. Wilson launches major upgrade of Wilsonweb. *Information Today*, 19(11), pp. 24, 49.
- Tenopir, C. (1994, March 1). Target, Freestyle, Win... : Searching takes on a new look. *Library Journal*, 119, 34.
- Tenopir, C., & Cahn, P. (1994). Target & Freestyle: Dialog and Mead join the relevance ranks. *Online*, 18(3), 31-47.
- Tomaiuolo, N. G., & Packer, J. (1998). Maximizing relevant retrieval. *Online*, 22(6), 57-65.
- Warner, A. (1990). Natural language processing: Current status for libraries. In F. W. Lancaster & L. C. Smith (Eds.), *Artificial intelligence and expert systems: Will they change the library?* (pp. 194-214). Urbana-Champaign, IL: University of Illinois.

Appendix A – Relevance Rating Form for Retrieved Citations

Query: _____

This query is (Circle one): Boolean Natural language

Citation #	Relevant? Y, N, or leave blank	Citation #	Relevant? Y, N, or leave blank	Citation #	Relevant? Y, N, or leave blank
1		18		35	
2		19		36	
3		20		37	
4		21		38	
5		22		39	
6		23		40	
7		24		41	
8		25		42	
9		26		43	
10		27		44	
11		28		45	
12		29		46	
13		30		47	
14		31		48	
15		32		49	
16		33		50	
17		34			

Appendix B – Information for Students

SJSU Students,

Do you have a class research project for that requires you to find journal articles in an online database? We can help!

Send us your research question, and we will get back to you with a citation list of 100 articles. All you have to do is review the list and tell us if any of the articles are relevant to your search. After you receive the list, we are also available to assist you in continuing your research in other databases.

This is part of a study we're conducting on the effectiveness of different document retrieval methods using WilsonWeb, one of the King Library's electronic databases. The purpose of this study is to determine how well the database responds to the information needs of undergraduate students.

Please contact me at lisango@slis.sjsu.edu for more details if you are interested in participating.

Thank you!

Sincerely,

Lisa Ngo

Graduate Student, School of Library and Information Science
San José State University
lisango@slis.sjsu.edu