17:610:530 – Principles of Searching

# Search for Anaphora: Report and Analysis of an Information Retrieval Experience

Michael J. Giarlo

School of Communication, Information and Library Studies

Rutgers, the State University of New Jersey

**Abstract**

This paper reports and analyzes all facets of an information retrieval (IR) experience, from the initial interactions with a user to the delivery of a final result set. Each facet is presented in sequential order and analyzed in terms of its relevance to the satisfaction of the user's information needs. Emphasis is placed upon the intermediary's searching experiences with the DIALOG, LexisNexis, Web search engines & the "invisible web," and digital libraries. The user for this particular paper was interested in current, scholarly information within the field of theoretical linguistics, and so DIALOG was by far the most useful source of information although Web search engines also provided some useful materials. The information retrieved primarily through DIALOG was judged to be ideal by the intermediary and was presented to the user who was satisfied with the results and eager to do further research.

**Introduction**

In order to satisfy a user's information needs, a number of decisions need to be made by the intermediary, some before a user can even be selected, some while engaging the user in the initial user-intermediary interaction or interview, and some must even be made "on the fly," in the middle of interactions with the IR system. The numerous decisions that the intermediary needs to make include selection of a user-intermediary interaction model, application of a general IR model, employment of specific searching strategies and tactics, identification and evaluation of information sources, delivery of results to the user, and the satisfaction of the user in terms of

whether the information need was handled in an effective, efficient manner. These decisions will be explained and justified herein.

While in many "real world" scenarios a user will seek out an intermediary for assistance, the assignment of which this paper is part was to play the role of an intermediary to seek out a user by polling for information needs. Before selection of a user may proceed however there is an initial question to answer, and that is "What model of interaction in information retrieval will be applied?" For the purposes of this assignment, the triadic model of interaction of information retrieval (Saracevic 1989) was clearly more appropriate than the dyadic model, since we are learning to become intermediaries, which are not included in the dyadic model.

**User modeling**

My initial attempts to locate a user with appropriate information needs failed, primarily due to the rather narrow audience I polled. After broadening the audience significantly, the information needs of one particular user, whom I would consider a "peripheral friend," stood out from the rest.

The user in question, whose name is John, is a 23-year old male living in the area of Boston, Massachusetts. John received his bachelor's degree in linguistics from Boston University two years ago and has recently been mulling over the possibility of further studying linguistics at the graduate level.

To that end, he would like to continue researching theories of anaphora, a phenomenon which greatly interested him during his undergraduate days. For the purposes of this paper, I will avoid going into a detailed explanation of anaphora. What is important is that I

also studied linguistics as an undergraduate, so I was a particularly well-prepared intermediary. The specific question in which John was interested is "What have the key developments been in theories of anaphora over the past five years?"

Considering John lives in Massachusetts and the tight time frame within which this project needed to be completed, I decided that I would set up a rather short, informal IR interview with John over the Internet. Before the interview itself, I decided to apply the "information problem detection" mindset to the mode of inquiry I would adopt. The other types of mindsets detailed are "query formulation process" and "database instructions," which did not seem as relevant to me as I was not concerned with the complexity of the user's search question nor the instruction of the user in terms of database usage. Information problem detection "is characterized by the intermediary's frequent use of elicitation related to the user's information problem to diagnose the user's real information need" (Wu 2003).

We conversed using one of the popular "instant messaging" networks for approximately thirty minutes, during which I asked a number of questions intending to determine what previously unforeseen limits the user might have had in mind. I discovered that John was not only seeking information on theories of anaphora since 2000, but that he also expected the paper to be written in English, which is not a trivial point in the field of linguistics where much of the research is done in foreign languages and many of the researchers are fluent in at least two languages. Additionally, John expressed very strongly his distaste of theories based on the work of Noam Chomsky, arguably the most brilliant and prolific linguistic theorist of our time, upon which much of modern linguistics is based. While I noted his distaste of Chomskyan theories, I felt I would be remiss in my duties as an

intermediary if I filtered out Chomsky-influenced theories should they have proven be to prominent during the past five years.

Considering that John has an excellent grasp on linguistic theory and a comprehensive knowledge gained from undergraduate study and that he is interested in research, it became obvious early on that the materials in which he was interested would come from scholarly journals. In fact, he indicated this himself. Ideally, the user wanted between five and ten references returned to him so that he could search out primary source materials and read them in his leisure, which made the task of organizing and collecting results much simpler and quicker for me.

**Search Task and Questions**

The user fortunately had a very clear idea of what his information needs were and was able to express them quite effectively to me. Given my similar background in linguistics and a pre-existing rapport, few questions arose during the stage of search task definition. As such, defining the search task was a trivial step: locate and return between five and ten references to developments in theories of anaphora published in English since the year 2000.

**Search Strategies and Tactics**

Before expounding upon the search strategies and tactics employed, I decided to follow the five basic phases and seven generic guidelines to perform a search (Zins 2000):

1. Assignment
   (1) Define the search assignment.
2. Resources
   (2) Locate the resources.

3. Search Words
   (3) Choose the search words.
4. Method
   (4) Select the proper search methodology.
   (5) Execute the search.
5. Evaluation
   (6) Evaluate the results.
   (7) If necessary, repeat the search by refining previous decisions

First and foremost, I needed to make some determinations on the relevant concepts for my search strategy. I decided that the key concept was "anaphora." This term is used rather uniquely to describe the phenomenon being studied. A synonym for the general phenomenon of anaphora is the term "anaphor," which is the basic unit of interest. Should these terms return an insufficient set of results, the user and intermediary were both comfortable deciding that this meant there were no records to be retrieved. That is, the concept of "anaphora" is central and unique for the user's information needs, so few if any synonyms need to be considered.

The search assignment has already been defined during the IR interview, so phase 1 had been completed. Phases 2 through 5 were run through at least once for each type of information source.

*DIALOG*

With the concepts already decided upon, I opened up the DialIndex bluesheet and did a visual scan of the files for appropriate databases. The "LANGUAGE" file appeared to be the most promising, so I chose that for my search, issuing the "b 411" and "sf language" commands.

My first search was "anaphor?", which appeared in ERIC, Social SciSearch, PsycINFO, Dissertation Abstracts Online, Gale Group Business A.R.T.S., British Education Index, Wilson Social Sciences Abstracts,

Information Science & Technology Abstracts, Wilson Humanities Abstracts Full Text, and Arts & Humanities Search. I decided to use all these databases and began using OneSearch to query them. Here are my queries and the number of results (after removing duplicates):

| | |
|---|---|
| PY=2000:2004 AND anaphora | 467 |
| PY=2000:2004 AND (theory or theories) AND anaphora | 258 |
| PY=2000:2004 AND theor?(2n)anaphora | 15 |
| PY=2000:2004 AND (theor? AND anaphora)/de | 4 |

The first two searches yielded far too many results to browse through so I instead displayed the last two which were much higher in terms of precision, looking through 19 results. Of these, 8 were quite relevant so I used DIALOG's "KEEP" command to save them.

From the last entry, three results suggested to me that a relatively new theory of anaphora had been developed, known as Neo-Gricean Pragmatic Theory. At this point, I decided to adopt something of a berrypicking model (Bates 1989), going back and issuing a new search based on these results. My new search on neo()gricean() pragmatic()theory yielded only 7 results and I kept the first result as it was the original paper on the subject.

*LexisNexis*

Building upon the success of the searches I executed on Dialog, I decided to stick with my original key concept of "anaphora." I started out with Nexis.com's Quick Search given its ability to search all sources at once. Since Lexis-Nexis isn't well-suited for retrieval of information about linguistic theory – the more scholarly, academic

sources made available by Dialog are far more appropriate, and yielded fruitful results – I figured that starting out with the broadest net possible would be the wisest course of action.

Additionally, I perused the subject categories at Nexis.com to see if any were appropriate to the subject of linguistics. Not surprisingly, none fit the bill.

At Lexis.com, I began with the "Find a Source" tab since none of the other tabs – Legal, News & Business, and Public Records – are applicable to theoretical linguistics. Unfortunately, I failed to find any sources for information about linguistics.

My Nexis.com QuickSearches on "anaphora," "pronominal," and "anaphor" all yielded no results. Even though I intended not to use any synonyms for "anaphora," I decided to broaden the net as much as possible on LexisNexis. Unfortunately, my searches on "reflexive," "reflexive pronoun," and "pronoun" retrieved a number of documents, none of which were relevant.

*Search Engines and the "Invisible Web"*

The sites that I judged to be of the most potentially use were Google, Yahoo!, Librarians' Index to the Internet, Vivisimo, Complete Planet, the Virtual Library, and the Internet Public Library, given their scope and organization. Generally speaking, the web was a fairly good source of information though not nearly as much as DIALOG. Most the materials found on the web were "less official" than those in the scholarly journals, consisting primary of conference papers and author-published pre-prints.

For the search engines – Google, Vivisimo, and Complete Planet – I started by doing a simple search on the word "anaphora," which is the key concept; if a document is relevant to the user's information needs, the word "anaphora" will appear in it. To narrow down the results if necessary, I issued a subsequent search on "theories of anaphora."

For the directories and virtual / digital libraries, I obviously needed to adopt a different strategy given the marked differences in the organization of these websites, starting out by looking for general information about linguistics and then drilling down.

Searching Google for "anaphora" returns over 68,000 results, so I narrowed down my search to "theories of anaphora," which netted 261 results. Google, unfortunately, only supports the following options for limiting by date: "past 3 months," "past 6 months," "past year," and anytime. Therefore, I was not able to specify that I want results from the past five years. Additionally, there is no information about the date of the documents in the result list. Therefore, I scanned the first 30 results to look for relevant documents.

Searching Vivisimo for "anaphora" returns a number of results, most of which are irrelevant judging by the way Vivisimo clusters results. Issuing the narrower "theories of anaphora" search returns fewer results as would be expected, and includes a cluster named "linguistics" with 10 documents.

Complete Planet allows for the selection of "Deep Web" databases, so I began by browsing their directory tree of subjects. First I selected Humanities, then Language & Linguistics, then I issued separate searches for "anaphora" and "anaphor," both of which retrieved 0 results. (Note, I actually checked in the Social Sciences category

first, as linguistics is widely considered a social science rather than a member of the humanities field.)

As Yahoo! has a similar directory structure to Complete Planet and LII, I first clicked on the Social Science category (trusting that Yahoo! had gotten right what Complete Planet botched). Beneath that category is the Linguistics and Human Languages subcategory. Within this subcategory, which contains several subcategories that could have relevant information, I ran searches for "anaphora" and "anaphor," both of which retrieved 0 results.

As I could not locate a category for linguistics in the Librarians' Index to the Internet, I instead ran a general search for the term "linguistics," which retrieved 36 results, all of which were too general. Subsequent searches on "anaphora" and "anaphor" retrieved 0 results.

VLib, the Virtual Library, also lists linguistics as a member of the humanities. However, when I clicked on the main linguistics link, a 404 File Not Found error resulted! I suppose this is to be expected every now and again when dealing with web sources.

The Internet Public Library (IPL) lists linguistics as a subcategory named "Language & Linguistics" beneath the Arts & Humanities category. Issuing searches on "anaphor" and "anaphora" within this subcategory yielded 0 results.

*Digital Libraries*

For my search, I chose to browse and search all the digital libraries in both Lesk's and Tefko's lists, which I judge to be too numerous to list herein. None seemed particularly relevant to

theoretical linguistics, though I thought many of them might have a gem or two hidden somewhere.

For each digital library, I first tried to issue the broadest search I could think of, which was "linguistics." This produced zero results for many of the digital libraries. On the off chance that searching on “linguistics” returned too many results, I searched on "anaphora." I could not find a single useful resource, probably since my topic is most relevant to scholarly journals. In fact, scholarly journals are the perfect, if not only, resource for my user's question.

Of those that returned results, I could not find a single useful resource. Those that returned many such results, I tried to limit using "anaphora" or "anaphor" which almost always returned zero results.

The other strategy I used was to navigate the sites' subject directory trees. By and large, if a site had a linguistics category, I could not find any information within on theories of anaphora.

Search steps were mostly similar from digital library to digital library, although there are no standards for layout or search engines, so there were little variations. Most sites seem to have been designed somewhat intuitively, however, so it was not difficult to find the search and browse interfaces.

**Evaluation of Information Sources**

*DIALOG*

The DIALOG system provides a robust command language, allowing for quite powerful and precise searches, in addition to a wide variety of subject-specific databases. Combined, they make a very useful resource, especially for pulling up academic or scholarly information.

With DialIndex and OneSearch, the process becomes even more powerful, allowing searches across multiple databases. Also, the different views of results, such as Full records, KWIC, etc., are very helpful and save searchers considerable time.

Unfortunately, DIALOG isn't without some drawbacks. On a few occasions, session errors occurred and I needed to log back into DIALOG. Though a bit inconvenient, I had been saving my search strings to a separate text file to document the process, so the errors did not impact the searching process very much at all. Also, the system can be rather slow when issuing OneSearch requests especially when trying to retrieve broad result sets. A major disadvantage of having a multi-database system where all databases do not conform to the same standards is that support for certain basic and additional indexes is not available in all databases. That is especially unhelpful when issuing OneSearch commands, since one needs to check multiple bluesheets before confirming that all the databases do or do not support the same basic or additional indexes.

The DIALOG interface is quite effective, providing command-level searching for more proficient users and a guided search for less proficient users who may be more comfortable with category browsing. The ability to view all your search sets at once in sequential order, and access previous commands are nice interface tools as well.

*LexisNexis*

Advantages of Lexis-Nexis include a lengthy listing of high-level categories, the addition of a Quick Search to aid less experienced searchers, a well-designed and functional interface, and the Power Search interface which supports a robust and powerful command language,

not dissimilar from Dialog's. These features are beneficial as they support the needs of a range of information seekers, from the inexperienced to the expert.

The only problem I have had with Lexis-Nexis is that its materials were not appropriate to my user's information needs, which I would not quite label a disadvantage of the system as a whole. The system provides sufficient functionality, along with extra features listed in 5a, and the interface is designed to please one's sense of aesthetics.

Since I have used Dialog much more than I have used Lexis-Nexis, I fear that my preference for Dialog is based solely on these experiences. Were we given more time with Lexis-Nexis, and were my user's information needs different, I would possibly prefer Lexis-Nexis. Lexis-Nexis's inclusion of subject category listings, a quick search, and a number of additional operators which seem quite useful could possibly have swayed me towards Lexis-Nexis.

*Search Engines and the "Invisible Web"*

The advantages of these web sites are by and large ease of use and availability of access. That is, they may be used by inexperienced users with no knowledge of search strategies, and they are available to the general public, unlike Dialog and Lexis-Nexis. The other major advantage of these sites is the breadth of the material they cover, which is literally millions upon millions of websites.

Unfortunately, the advantages of these sites are inherently tied to their disadvantages. On one hand, there is an awful lot of information to weed through to get to relevant documents, but on the other hand, a great deal of the results are completely irrelevant.

Precision tends to be abysmally low with web searches, given the lack of authority control and the general ability of anyone to put anything on the web. Additionally, many of the search engines do not provide very robust features for searching. The tools are rather rudimentary, on the whole.

Since my user's information needs are related to scholarly material, I prefer using more traditional IR systems such as DIALOG and LexisNexis. Not only can I be sure that the databases I search are relevant but I can search them with a high degree of precision, thanks to the powerful command languages provided by Dialog and Lexis-Nexis. In general, though, I do prefer using web searches for finding information though that may be due to the fact that I am most often searching for non-scholarly information.

*Digital Libraries*

The main disadvantage of digital libraries has been subject coverage. That is, material that is germane to my user's topic is found exclusively in scholarly journals. Granted, some of this material is online in electronic journals provided by vendors such as Academic Search Premier and Ingenta, and other material is online in pre-print form. I do not consider these to be "digital libraries," however, since I can and have accessed the same information via Dialog. Digital libraries, in my opinion, serve all or most of the functions that traditional libraries do. OPACs and e-journals, then, would be major parts of a digital library, representing catalogs and periodicals in the traditional library world, but are not digital libraries in and of themselves.

As I stated above, scholarly journals are perfect for my user's topic, hence Dialog was the ideal IR system. Lexis-Nexis and digital libraries did not include relevant subject coverage for the most part. The web yielded a few hits, thanks for pre-prints and copies put online by authors and archivists, but finding the information was hit-or-miss. For my specific search task, Dialog was the perfect system. I don't have an objective preference for any particular IR system; it depends entirely on the subject matter for which I am searching, which in this case was very narrow and academic.

**Discussion**

Overall, I would rate my experiences with information retrieval as successful and informative. Not only have I been able to learn how to use numerous online IR systems, but I have been exposed to theories of different interaction models, various types of searching behavior, and a number of options regarding display of results to return to the user. I feel as though my topic was perhaps too narrow, however, as I received disproportionate experience with DIALOG and search engines as I did with LexisNexis and digital libraries. Within DIALOG, I had the most success using DialIndex and OneSearch, and rather than using the supplied controlled vocabulary, I chose rather to construct detailed phrase searches using the proximity operators.

Despite the fact that all of my results came from DIALOG and a couple general web searches, I am pleased knowing that my user, John, is satisfied with the information I was able to retrieve and intends to act upon it by searching for the materials himself and doing more exhaustive research.

**Conclusion**

The information retrieval experience as a whole contains a series of questions that must be addressed by the intermediary, whether through selection of appropriate models or by direct interaction with the user. After I selected the models I considered most appropriate for this assignment and conducted the initial user interview, I was able to explore the major search systems which was in itself a learning experience.

Having some knowledge and interest in the user's topic was of great use to me, providing a little extra motivation and background when issuing and evaluating searches. After all the questions have been answered and results have been evaluated, an intermediary such as myself can only hope that he has done his best in acting as the "middle man" between the cognitive needs of the user and the vast store of information residing on discrete systems, each with its own functionality and quirks.

**References**

Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13 (5), 407-424. O

Ellis, D., Wilson, T. D., Ford, N., Foster, A., Lam, H. M., Burton, R., & Spink, A. (2002). Information seeking and mediated searching. Part 5. User-intermediary interaction. Journal of the American Society for Information Science and Technology, 53 (11), 883-893.

Saracevic, T. (1989). Modeling and measuring user-intermediary-computer interaction in online searching: Design of a study. Proceedings of the Annual Meeting of the American Society for Information Science, 26, 75-80.

Wu, Mei_Mei, & Liu, Ying_Hsang (2003). Intermediary's information seeking, inquiring minds, and elicitation styles. Journal of the American Society for Information Science and Technology, 54(12), 1117-1133.

Zins, C. (2000). Success, a structured search strategy: Rationale, principles, and implications. Journal of the American Society for Information Science, 51 (13), 1232-1247.

**Appendix I – Results Returned to User**

The results I returned to my user came in two forms: a list of author & title references for either online retrieval or for acquisition through a local or academic library, and a list of URLs for web-based retrieval. The results are listed below for the sake of completeness.

*DIALOG*

Title: A Neo-Gricean Pragmatic Theory of Anaphora
Author: Yan, Huang

Title: Testing the Neo-Gricean Pragmatic Theory of Anaphora: The influence of consistency constraints on interpretations of coreference in Spanish
Author: Blackwell, SE

Title: Acquisition of binding of English reflexives by Turkish L2 learners: A Neo-Gricean pragmatic account

Author: Demiri, M

Title: Anaphora interpretations in Spanish utterances and the Neo-Gricean Pragmatic Theory

Author: Blackwell, SE

Title: Implementing the binding and accommodation theory for anaphora resolution and presupposition projection

Author: Bos, J

Title: Anaphora: lexico-textual structure, or means for utterance integration within a discourse? A critique of the functional-grammar account

Author: Cornish, F

Title: Discourse anaphora: Four theoretical models

Author: Yan, Huang

Title: Mental models and the interpretation of anaphora

Author: Garnham, Alan

*Search Engines and the "Invisible Web"*

Title: The Syntax of Anaphora (abstract)

Author: Safir, Ken

URL: http://ruccs.rutgers.edu/~safir/soa-abs.pdf

Title: The Syntax of Anaphora (full-text)

Author: Safir, Ken

URL: http://ruccs.rutgers.edu/~safir/soa-ms.pdf

Title: The Logic of Anaphora Resolution

Author: Beaver, David

URL: http://montague.stanford.edu/~dib/Publications/ac99paper1.pdf

Title: Form and Meaning in a Competitive Theory of Anaphora (abstract)

Author: Safir, Ken

URL:http://www.linguistics.ubc.ca/PRON/abstracts/safir.pdf

Title: Toward a Feature-Movement Theory of Long-Distance Anaphora

Author: Richards, Norvin

URL:http://minimalism.linguistics.arizona.edu/AMSA/PDF/AMSA-32-0900.pdf

Title: Presupposition or Abstract Object Anaphora? Constraints on Choice of Factive Complements in Spoken Discourse

Author: Spenader, Jennifer

URL:http://www.coli.uni-sb.de/~korbay/esslli01-wsh/Proceedings/19-Spenader.pdf