A LINGUISITICAL ANALYSIS OF WORLD WIDE WEB QUERIES

Bernard J. Jansen
Computer Science Program
University of Maryland (Asian Division)
Seoul, 140-022 Korea
Email: jjansen@acm.org

Amanda Spink
School of Information Sciences and Technology
The Pennsylvania State University
511 Rider I Building, 120 S. Burrowes St.
University Park PA 16801
Tel: (814) 865-4454 Fax: (814) 865-5604
E-mail: spink@ist.psu.edu

MAJ Anthony Pfiff
Department of English, United States Military Academy
West Point, New York 10996 USA
Email: pfiff@exmail.usma.edu

Please Cite: Jansen, B. J., Spink, A., & Pfaff, A. 2000. Linguistic Aspects of Web Queries. American Society of Information Science 2000. Chicago, November 13-16 2000.

See Other Publications

ABSTRACT

Terms are the basic building block of information retrieval (IR) queries. Queries are the primary means of the translating the user's information needs in a way that the IR system can understand. As such, terms and how they are used in queries are the essential components of a user's problem solving and decision making interaction with any IR system. If the terms, their semantics, and the query syntax in which they are used could be modeled, the IR system could be tailored or to this model, thereby providing greater assistance to the user in finding relevant information. In pursue of this goal, we analyzed at three levels a transaction log containing 51,473 queries posed by 18,113 users of *Excite*, a major Internet search service. We extracted the terms and examine their rank and frequency distribution compared with large collections of English documents. We then examine individual queries to isolate query structure syntactic patterns. Finally, we focus on the entire series of queries from a user. Based on these three levels of analysis, we were able to develop a linguistic model to classify queries into five (5) general categories. We discuss how these finding and the linguistic model relate to linguistic theory. We conclude with the implications of this user model on system design of IR systems.

INTRODUCTION

Information retrieval (IR) and Web user modeling is a growing area of research as the realization has increased that the user must be considered as part of the complete IR system (Brajnik 1987; Saracevic, Spink, and Wu 1997). Saracevic, Spink, and Wu (1997) reviewed the history and current state of user modeling research in traditional IR systems. There is also a growing body of literature focusing on IR in the context of the Web (Croft, 1995; Jansen, Spink, & Saracevic, 1998a,b; Jones, Cunningham, & McNab, 1998; Lawrence & Giles, 1998; Lynch, 1997). However, many Web studies have focused on user characteristics and empirical analysis of users' queries, with little attention to theory development or theory application.

In this study, we investigate the applicability of linguistic theory and linguistic analysis of user queries to the improvement of IR and Web system. Users of such systems are natural language users. Knowing how natural language users structure their queries in an attempt to model their information need may reduce the gap between how a computer works and how the "typical user", (i.e., a user with limited knowledge about how an IR system works) thinks the system does work. By analyzing the user queries for structure, syntax, and semantics, we may be able to develop strategies that will benefit IR system design. In pursuit of this line of investigation, we analyzed a transaction log from the *Excite* search engine, a major Web media company. This paper reports the methods, findings and results from a linguistic analysis of this corpus of queries from users of the *Excite* search engine. The next section of the paper discusses in detail the data corpus used in this study.

EXCITE DATA CORPUS

Founded in 1994, *Excite,* Inc. is a major Internet media public company that offers free Web searching and a variety of other services. The company and its services are described at its Web site [http://www.excite.com]. Only the search capabilities relevant to our results are summarized in this paper. *Excite* searches are based on the exact terms that a user enters in the query. Capitalization is disregarded, with the exception of logical commands AND, OR, and AND NOT. Stemming is not available. An online thesaurus and concept linking method called Intelligent Concept Extraction (ICE) is used, to find related terms in addition to terms entered. Search results are provided in a ranked relevance order. A number of advanced search features are available. Those that pertain to our study are described here:

- As to *search logic, Boolean operators* AND, OR, AND NOT, and parentheses can be used, but these operators must appear in ALL CAPS and with a space on each side. When using Boolean operators ICE (concept-based search mechanism) is turned off.
- A set of terms enclosed in *quotation marks* (no space between quotation marks and terms) returns answers with the terms as a phrase in exact order.

- *A + (plus) sign* before a term (no space) requires that the term must be in an answer.
- *A – (minus) sign* before a term (no space) requires that the term must NOT be in an answer. We denote plus and minus signs, and quotation marks as modifiers.
- A page of search results contains ten answers at a time ranked as to relevance. For each site provided is the title, URL (Web site address), and a summary of its contents. Results can also be displayed by site and titles only. A user can click on the title to go to the Web site. A user can also click for the next page of ten answers.
- In addition, there is a hypertext option, *More Like This*, which is a relevance feedback mechanism to find similar sites. When *More Like This* is clicked, *Excite* enter this in the transaction log as a query with zero terms.

Each record in the transaction log contained three fields. With these three fields, we were able to locate a user's initial query and recreate the chronological series of actions by each user in a session:

- *Time of Day:* measured in hours, minutes, and seconds from midnight of 9 March 1997.
- *User Identification:* an anonymous user code assigned by the *Excite* server.
- *Query Terms:* exactly as entered by the given user.

Focusing on our three levels of analysis, sessions, queries, and terms, we defined the variables in the following way.

- *Session:* A session is the entire series of queries by a user over time. A session could be as short as one query or contain many queries.
- *Query:* A query consists of one or more search terms, and possibly includes logical operators and modifiers.
- *Term:* A term is any unbroken string of characters (i.e. a series of characters with no space between any of the characters). The characters in terms included everything – letters, numbers, and symbols. Terms were words, abbreviations, numbers, symbols, URLs, and any combination thereof. We counted logical operators in capitals as terms, however, in a separate analysis we isolated them as commands, not terms.

Some general statistics about the data corpus are presented in Table 1.

No. of users	No. of queries	Non-unique terms	Mean of terms And Range	Unique terms with case sensitive	Unique terms without case sensitive
18,113	51,473	113,776	2.21 0-10	27,459	21,837

Table 1. Numbers of users, queries, and terms

As one can see, there were over 18,113 users and 51,473 queries. So, it was a large number of users and queries, and therefore, a very rich data corpus. The next section of the paper discusses the term, query and session analyzes conducted to form the basic for the linguistic analysis.

TERM LEVEL OF ANALYSIS

We first focused on the term level of analysis. We separated the queries into terms. A term was any series of characters bounded by white space. There were 113,793 terms (all terms from all queries). After eliminating duplicate terms, there were 21,862 unique terms that were non-case sensitive (in other words, all upper cases are here reduced to lower case). In this distribution logical operators AND, OR, NOT were also treated as terms, because they were used not only as operators but also as conjunctions. We discuss terms from the perspective of their occurrence and their fit with known distributions.

Occurrences

We constructed a complete rank-frequency table for all 113,793 terms. The number one ranked terms occurred the most frequent, the second ranked term, occurred the second most frequent, etc. Out of the complete rank-frequency-table we took the top used terms i.e. those that appeared 100 times or more, as presented in Table 2.

Term	Frequency	Term	Frequency	Term	Frequency
and (incl. ‘AND’, & ‘And’)	4828	&	188	estate	123
Of	1266	stories	186	magazine	123
The	791	p****	182	computer	122
Sex	763	college	180	news	121
Nude	647	naked	180	texas	119
Free	610	adult	179	games	118
In	593	state	176	war	117
Pictures	457	big	170	john	115
For	340	basketball	166	de	113
New	334	men	163	internet	111
+	330	employment	157	car	110
University	291	school	156	wrestling	110
Women	262	jobs	155	high	109
Chat	256	american	153	company	108

On	252	real	153	florida	108
Gay	234	world	152	business	107
Girls	223	black	150	service	106
Xxx	222	porn	147	video	105
To	218	photos	142	anal	104
Or	213	york	140	erotic	104
Music	209	A	132	stock	102
Software	204	Young	132	art	101
Pics	202	History	131	city	100
Ncaa	201	Page	131	porno	100
Home	196	Celebrities	129		

Table 2: Listing of Terms Occurring More Than 100 Times (**** = expletive).

An interesting aspect of the rank – frequency table was the distribution of terms. There were 74 terms (of the 21,837 unique terms) that occurred more than 100 times in all queries. On the other end of the spectrum**,** there were 9,790 terms that occurred only once. The 74 terms that were used 100 or more times had a frequency of 20,698 appearances as search terms in all queries. They represent 0.34 % of all unique terms, yet they account for 18.2 % of all 113,776 search terms in all queries. If we delete the 11 common terms that do not carry any content by themselves (*and, of, the, in, for, +, on, to, or, &, a*) that altogether had 9,121 occurrences, we are left with 63 subject terms that have a frequency of 11,577 occurrences – that is 0.29% of unique subject terms account for 10.3% of all terms in all queries. Interestingly, the high appearance of '+' represents also a probable mistake – the inclusion of space between the sign and a term, as required by *Excite* rules. Similarly, '&" was used often as a part of an abbreviation, such as in *AT&T*, but also as a substitute for logical AND, as in *ontario & map.* In the latter case, it is a mistake and would appear as a separate term.

On the other end of the distribution, the 9,790 terms that appeared only once amounted to 44.78% of all unique terms and 8.6% of all terms in all queries. The tail end of unique terms is very long and warrants in itself a linguistic investigation. However, we could find no comprehension studies of what terms, the distribution of those terms, the modification of those terms, etc. of Web queries.

Distribution of Terms

Since we could find no previous work in this area, we decided to determine if the rank – frequency distribution of terms fit known distributions. Based on well-known work with large collections of English text, it would be reasonable to assume that the rank – frequency plot would fit a Zipf distribution. A very brief introduction to Zipf's Law follows.

Zipf's Law

Zipf's law is the observation that frequency of occurrence of some event as a function of the rank, when the rank is determined by the above frequency of occurrence, is a power-law equation. The most famous example of Zipf's law is the frequency of English words. If the terms in a collection are ranked (r) by their frequency (f), they roughly fit the relation r_t * f_t = C, which is known as "Zipf's law". Different collections have different constants C, but in English text, C tends to be about N / 10, where N is the number of words in the collection. When these rank – frequency equations are plotted on a double log graph (i.e., the log of rank by the log of frequency), there is a linear relationship with a slope of negative one.

Rank – Frequency Plot of Query Term Distribution

We constructed a graph of rank – frequency distribution of all terms. This graph is shown in Figure 1.

Figure 1: Rank vs. Frequency (log) of All Terms.

The straight line is close to what one would expect if the rank – frequency of terms adhered to the Zipf distribution. However, the resulting distribution of the actual data (the non-straight line) does not behave as expected. The distribution of data seems to be unbalanced at ends of the graph, the high and low ranking terms. The data does not conform to the Zipf distribution until about a rank of 1,000 (i.e., Rank (log) = 3), that is the term that ranked at approximately 1,000 in the rank – frequency table. At the beginning, the distribution falls off very gently, and toward the end it shows discontinuities (i.e., plateaus) and an unusually long tail, representing terms with frequency of one. What this distribution and comparison with the Zipf distribution would seem to indicate is that users have one technique, or maybe language, to compose documents and another to compose queries.

QUERY LEVEL OF ANALYSIS

At the term level, there seemed to be indications of deviation between the language of queries and the language in which people talk and write. We therefore decided to examine the data at the query level. Specifically, we were interested in the length (i.e., the number of terms) of the queries. This information is displayed in Table 3.

Terms in query	Number of queries	Percent of all queries

10	185	0.36
9	125	0.24
8	224	0.44
7	484	0.94
6	617	1
5	2,158	4
4	3,789	7
3	9,242	18
2	16,191	31
1	15,854	31
0	2,584	5

Table 3: Number of terms in queries. (N queries = 51,473).

On the average, a query contained 2.21 terms. Table 3 shows the ranking of all queries by number of terms. Percent is the percentage of queries containing that number of terms relative to the total number of queries. Web queries are short. About 62% of all queries were one or two terms. Less than 4% of the queries had more than 6 terms. This is substantially lower than people's average English utterances or written sentences. For the example, the average sentence length for this paper is approximately nineteen (19) words. Similar to the deviation from the expected term distribution, the short exchanges (i.e., queries) between the user and the computer would seem to indicate that there is something different about this communication, compare to human – human communication.

SESSION LEVEL OF ANALYSIS

Users can send repeated queries to the computer. This entire sequence of queries by a users is called a *session*. We examined the session level to gain further insight into this user – computer dialogue. We first classified the 51,474 queries as to *unique, modified,* or *identical* as shown in Table 4.

Query Type	Number	Percent of all queries

Unique	18,098	35
Modified	11,249	22
Identical	22,127	43

Table 4: Unique, Modified, and Identical Queries.

A unique query was the first query by a user (this represents the number of users, including an error). A modified query is a subsequent query in succession (second, third …) by the same user with terms added to, removed from, or both added to and removed from the unique query. Unique and modified queries together represent those queries where user did something with terms. Identical queries are queries by the same user that are identical to the query previous to it. They can come about in two ways. The first possibility is that the user retyped the query. Studies have shown that users do this (Peters, 1997). The second possibility is that the query was generated by *Excite*. When a user views the second and further pages (i.e., a page is a group of 10 results) with the same query, Excite provides another query, but a query that is identical to the preceding one. The unique plus modified queries (where users actively entered or modified terms) amounted to 29,437 queries or 57% of all queries. If we assume that all identical queries were generated as request for viewing subsequent pages, then 43% of queries come as a result of desire to view more pages after the first one.

We were interested in the length of user sessions, that is the number of queries per session. The results are displayed in Table 5.

Queries per user	Number of users	Percent of users	Queries per user	Number of users	Percent of users
1	12,068	67	10	17	0.09
2	3,501	19	11	7	0.04
3	1,321	7	12	8	0.04
4	583	3	13	15	0.08
5	287	1.6	14	2	0.01
6	144	0.80	15	2	0.01
7	79	0.44	17	1	0.01
8	32	0.18	25	1	0.01

9	36	0.20			

Table 5: Number of Queries Per User.

Most users used only one query in their session, others used a number of successive queries. The average session, including all three query types, was 2.84 queries per session. This means that a number of users went on to either modify their query, view subsequent results, or both. The average session length, ignoring identical queries, was 1.6 queries per user. Table 5 includes only the 29,337 unique and modified queries. We ignored the identical queries in order to concentrate only on those queries where users themselves did something to the queries. A big majority of users did not go beyond their first and only query. Some 67% of users had one and only query. Query modification was not a strong trend. Similar to the query level, it appears that users provide the computer very limited clues about their information need. With the average query at about two (2) terms and the average session at about two (2) queries, the user is only providing the computer about four (4) terms during the entire dialogue.

The quantitative analysis would seem to indicate that there might be a substantial difference between how people communication with each other and how they communicate with a computer. We were interested in the applicability of a linguistics model of communication to the term, query and session level analysis.

LINGUISTIC ANALYSIS

When people communicate with each other, the hearer/reader tries to comprehend what the speaker/writer is communicating by observing the syntax of the sentence, the semantics of the words, and how they affect each other. What a word means depends, in part, on its lexical category (i.e., noun, adjective, verb, etc.). Where words, belonging to a particular lexical category, go in a particular expression depend on the syntax of the language in use. In English, the modifying word almost always precedes the word that it modifies, as in the expression "red chair." For example, the word "beautiful" is an adjective. When we hear it, we expect it to always precede the word it modifies. In fact, it would sound odd if it went after the word it modifies, as in "women beautiful." (This was an actual query from the data set.)

Sometimes, however, it is not clear to what lexical category a word belongs. Consider the expression "soccer team", which was also an actual query from the data set. Which word modifies which? The answer cannot be determined by looking at the form of the words (as one could with the modifier "beautiful"), but only by where the words go in the expression. Because in English syntax, the modifying word precedes the word that is modified, we know that ‘soccer’ modifies team. When a noun, like "soccer" modifies another noun (in this case "team") it becomes an attributive noun. In short, attributive nouns function like adjectives, but do not have the form of an adjective. In this way, the syntax of language projects onto the semantics of the expressions allowed by the syntax. With this linguistic base, we now move to results of the lexical analysis.

LEXICAL ANALYSIS

For the purposes of this preliminary work, we performed a lexical analysis of the first 511 queries from the data set. We examined the lexical patterns for individual queries as well as for entire sessions. The queries examined all appear to use English terms. While a complete analysis will require the examination of a much larger set, some interesting results emerged from analysis. Generally, one can say that users do not apply the normal rules of English syntax in any coherent or consistent manner. This is in line with our expectations following our term analysis. Users rely on a variety of lexical patterns to "explain" (i.e., formulate the query) to the "computer" (i.e., the IR system) what information, item, or topic they were trying to locate. Even in sessions where users performed multiple queries, the query patterns often vary widely and seldom conform to the rules of English syntax. From a linguistic point of view, there is no "language" to Web queries. A language must have some rules of syntax that permit one to distinguish a well-formed from an ill-formed phrase. There does not appear to be any such syntax with web queries. A discussion of linguistic theory is presented later in the paper.

While there did not seem to be any grammatical consistency to the queries, the syntax of the queries did seem to fall into five rather broad categories. The five categories are listed below, followed by a discussion of each. Later, we will explore the implications of these patterns.

- Adjective and noun phrases where one word was modified and where the others were doing the modifying.
- Complete and grammatically correct English sentences.
- Phrases comprised of verbs or verbals.
- Random strings of words of a variety of lexical categories but which seem to belong to the same category.
- Miscellaneous.

Adjective and Noun Phrase

This first category was by far the most represented, 458 of the 511 queries. Most of the queries in this category conformed to normal English syntax where the modified word (usually a noun, N) is the last word on the right, and the modifying words (usually an adjective ADJ or AN) are to the left. Additionally, the least restrictive modifier is closest to the modified word and the most restrictive modifier is farthest away. For example, in the query "brazillian soccer teams" (sic), the terms "brazillian" and "soccer" modify the term "teams". The term "brazillian" is the more restrictive relative to the modifier "soccer." When a noun, like ‘soccer’ modifies another noun (in this case "team") it becomes an attributive noun. In short, attributive nouns function like adjectives, but they do not have the form of an adjective.

In some cases , the noun being modified came first, as in the query "women beautiful." In this case, the user begins with the broadest category and then seeks to modify it into a more specific category. This is like shopping in a department store. You first ask the

doorman where is the shoe department, then ask the department clerk where are the running shoes then (based on assumptions you are making about who you are talking to) and then ask the sales rep where are the Nikes and so on. Here the user begins with the broadest category and then seeks to modify it into a more specific category. Later, we explore the implications of these patterns.

Grammatically Correct

In regards to the second category (14 of 511 queries), almost all queries of this type took the form of a question. Further, almost all took the form of a Wh- phrase. A Wh-phrase is an interrogative phrase that begins with words like what, where, when, how, why, which, and whose. A typical query of this type is: 'what is empty space in the universe composed of?' In nearly all of these sentences, the verb almost always had a two-place argument structure, which were usually theta marked as agent and theme or agent and location. This theta-marking pattern is also true of those few phrases that contained a verb.

Theta-marking is a way of delineating what kinds of words can be used as arguments for a particular verb. For instance, the verb *kill* has a two-place argument structure (e.g. The boy killed the deer). This is usually formally represented Kbb, where K represents the predicate *kill* and the b represents the boy and the d represents the deer. But not just anything can go in those places. For the verb *kill* one of the arguments must be something that can kill and the other something that can be killed. We can call the first the **agent** and the latter the **patient.**

We should note that there is not general agreement among linguists regarding what how many thematic roles there are or what their labels may be. But what is important is that the thematic category is going to limit the lexical category of possible responses. For example, in the case of an agent, it will almost always be a noun phrase such as, "The boy." This means that in the event a word can have more than one lexical category (for example, "play," it can be a verb as well as a noun), knowing the theta-marking of a particular verb will determine which lexical category the word falls in. Additionally, theta-marking imparts some semantic information about the word. For example, an agent not only is almost always a noun phrase, it also has to be something capable of causing an effect (in this case, death). Additionally, the patient must be something capable of receiving an effect (again, in this case, death).

Verbal Phrase

Th category (11 of 511 queries) were queries that contained verbs or verbals (i.e., a noun that had –ing added to it and which functions as a participle and/or a gerund) but which were not complete, grammatically correct English sentences. This was by far the most under-represented category. The queries containing verbals outnumbered the queries containing verbs six (6) to five (5). In many cases, the verbals stood alone, making it impossible to determine if they were meant as gerunds or participles, (e.g. as with the query 'hunting'). Where it was possible to determine, we discovered that most verbals

were gerunds. In this category, most of the verbs (including the root verbs the verbals were created from) had a two-place argument structure and were theta marked for agent and theme or agent and location. The ones that had only a one-place argument structure were theta marked as agent. A typical example of a verb phrase query was "boy and wolf cried", and an example of a verbal phrase query was "flood plains flooding."

Random Category

The fourth category (13 of 511 queries) contains those expressions that contained a series of words of varying lexical categories and which seem to defy syntactical categorization. The query "'alicia silverstone' cutest crush batgirl babysitter clueless" serves as a good (and one of the few non-x-rated examples) of this particular pattern. In this case it is not clear at all that the words are serving in the syntactic capacity that one would expect from their position in the query. This pattern does not conform to a standard, grammatically correct English sentence or phrase nor does it seem to conform to the first pattern we analyzed where one word is modified and the others do the modifying. So, while we can isolate out the lexical categories of most of the words, this does not help one make sense of the expression. It is also significant that one cannot pick out the lexical category of all the words, for example: "crush." Since the expression does not conform to a Standard English syntactical pattern, one can not tell if the word is a noun (as in "I have a crush on her") or a verb (as in "I will crush you").

While there does not seem to be a syntactic account for the meaning of this expression, there is a semantic one. The terms all seem to relate to a particular movie actress. A human (with the appropriate background) can tell this because each one of these words has something to do with the actress Alicia Silverstone and the movies she makes and roles that she plays. This will have some interesting implications later on when we offer some strategies for handling this kind of query.

Miscellaneous

We have included in the miscellaneous category (15 of 511 queries) any query pattern represented less than ten (10) times. The most prevalent of these are queries concerning URLs, email addresses, and grammatically incorrect English phrases, most being proper names. There were nine (9) URL and one (1) email address and five (5) queries that contained prepositions that were not grammatically correct English sentences. Most seemed to be associated with a proper name such as 'university of otago'. Since this category is of little interest to a linguistic analysis, we will not include them in the discussion section.

DISCUSSION

While we can group the lexical patterns into categories, it remarkable that there were so many different patterns and that so many queries did not conform to the basic syntax of English. This is an important point. We are not talking about users making simple mistakes in syntax, for which our high school grammarians would take points off. The

deviation from English syntax was much greater than simple a comma splice or run-on sentences. What this data indicates linguistically is that users are abandoning the way they think and communicate in English in order to communicate with the computer.

One explanation for this may be that as human users interact with the computer, they find that the syntax they normally relied on for effective communication did not have the effect that it normally had in a conversation with other humans. For instance, one grammatically correct query was: ‘what is the measurement and area of a one gallon can?’. We submitted this query to a major Web search engine on October 30, 1998 at 1723 and received 2,749,887 results, the first ten of which did not contain relevant information. Given performance such as this, users may realize that communicating with the computer the way they would with another human does not get the information they want, therefore, they change their communication strategies.

Ignoring the Miscellaneous category, since it contains no linguistic interest, it appears that user’s communication strategies can be classified in one of the four (4) categories listed. At least these categories may provide a starting point for describing those strategies. Given the overwhelming number of queries that fall under the first pattern, *Adjective and Noun Phrases*, it seems that this particular strategy either works best or is the default for many human users when they are not sure what syntax applies.

IMPLICATIONS FOR SYSTEMS DESIGN

Several aspects the findings have implications for system design in Web and possibly information retrieval in general. From the above discussion, at least three strategies for system design emerge for addressing the lack of syntax and problems surrounding user’s implementation of Boolean logic. Web and IR Systems could "recognize" certain syntactical patterns like those described above. For example, let us look at the *Adjective and Noun Phrases*, where the modified word is last in the series and the modifying words precede it, AN/ADJ_1 AN/ADJ_n N. While this is a simple pattern, it is rich in information. Just by its form, one knows which word contains the category of information the user is seeking. One also knows, of the modifying words, which is most and which is the least restrictive. A computer can perform this simple evaluation. n instances where there is a verb, the *Verbal Phrase* category, if the Web or IR system can detect the theta-structure of the verb, it will "know" what kind of item to look for, even if the system cannot tell to what category the item belongs. For the Random Category, a thesaurus of terms based on some stored dictionary or perhaps collaborative thesaurus based on previous searches could suggest categories to the system. For example, if queries from previous users contained terms such as: "batgirl babysitter clueless" along with "alicia silverstone", the IR system could categorize these terms. In fact, this is similar to how the *Excite* on-line thesaurus works, except *Excite* uses these as terms to suggest to the users.

LINGUSITIC THEORY

Most of the discussion so far has revolved around how our understanding of how particular syntax determines, in part, how one understands the meaning of an expression.

However, expressions are composed of words and until computers can understand how human beings convey meaning with words, computers will always be limited in how they interact with humans. Even if a computer contained an exhaustive dictionary that compiles the list of all possible meanings of a word, it still would not help the computer choose between meanings nor would it help the computer understand when new meanings are generated. While it is not possible to offer an exhaustive account of how this works in humans, we will try to do characterize, in some small degree, how human beings go about understanding the meaning of words. We can use this as a model to get an idea of the current limits of present day computer.

People understand the meaning of a word via a four part explanatory scheme. As Moravcsik (1990) states, a word has meaning in virtue of which element of reality counts as that word. Additionally, the meaning of a word is conveyed by up to four factors. These factors are the m-factor, the s-factor, the f-factor, and the a-factor. The m-factor applies to the constituent parts of a thing, everything from material substances to events and to abstract entities like arguments. Knowing the m-factor within the meaning of a word allows one to place the members of the extension in the correct category. These categories are (a) abstract, (b) material entity, and (c) event or state (d) objects of senses (e) transcategorial and (f) modifying elements. We can take any descriptive word or phrase, and run it through this list, and locate its m-factor".

The s-factor differentiates different kinds within the same category. These kinds are differentiated partly in terms of individuation and persistence, and partly in terms of qualitative difference. Thus, to determine the s-factor for a word involves answering the following questions in regard to a particular word:

What if any principle of individuation is tied to it?

What if any principle of persistence is tied to it?

What are the qualitative conditions that, given ordinary linguistic competence, will separate fully or partially, the items in the extension of the word from other items belonging to the class with the same m-factor, but falling within different extensions?.

While some words will have only constitutive and structural m- and s-factors), others may also have agential and functional (a- and f-) factors. Much in the same way it worked for Aristotle, the a-factor includes the necessary causal properties of items that fall within the range of things the word can denote. Similarly, the f-factor includes the functional properties of the word that are necessary to understand it; for example, under normal conditions a pen is used for writing, but not for stabbing someone - even though it could conceivably serve that function. The a-factor ranges over the causal properties that are the parts of the meanings of some words. It captures the fact that in the explanation of what some things are, their origin or causal potency is an essential ingredient. Things like artifacts and action words that require certain types of entities all have a-factors. For artifacts, the a-factor lies in who made it and who uses it. Action words such as ‘walk’ or ‘write’ have a-factors in relation to the kinds of entities that do them and the since what

counts as falling under them must have certain effects. The a-factor in walking lies in the fact that some types of locomotion resembling the paradigm of walking counts as such and others do not. A word like ‘writing’ has an a-factor in that writing means producing a set of symbols.

If it is not clear whether or not a word has an f-factor, one may ask, "Could this word have any meaning or range of application in a universe in which there can be no purpose, aim, function or result?" Thus a word like ‘number’ has no f-factor, but a word like ‘carpenter’ does. If all the meaning in a word meaning is conveyed this way, it can be represented in a general form as R (m, s, a, f) where R stands for the relation that ties the four factors together. If we take these four schemes as the base of our system of categories what we notice is that when words take on different meanings these factors change. As these factors change often the category of thing something belongs to changes. We also notice if we change too many of these factors the word ceases to have a recognizable meaning (give example). This is much like Neurath’s boat where a boat at sea needs repairs but cannot put into port. If there is a hole in hull, the crew takes the mast and cuts it into planks to patch the hole. But now the boat cannot go anywhere. So the crew dismantles the rudder to make a mast, but now it cannot steer itself to port. The crew is then faced with a difficult choice. Either float aimlessly and forever or be able to move and steer, but sink. The meaning of words work in much the same way. Just as if the crew chose to change too much on the boat to get it to move, it would no longer be a boat. Similarly, if we change the content of too many of the factors that convey a word’s meaning, at the same time, it ceases to have a recognizable meaning.

Thus, the more things we have to keep it "afloat" the more things we can be "repairing," that is changing so that new meaning can be generated. This suggests that there is a finite number of different meanings for a word at any given time. This also suggests that the relevant categories a word can belong to are finite.

CONCLUSION

Web and IR systems currently model the user’s information need via the query. However, most Web and traditional search IR engines follow a statistically query term and document term comparison. The premise of this analysis is that if one can correctly model the query, it would be a major step forward in correctly modeling a user’s information need. Previous IR modeling has focused on the user – system discourse, not on the query. Is there a linguist component to IR research? Is there a linguistic identification for query structure? It appears that there is some basic syntactic structure. User modeling must take into account the syntax and semantic of the query. Syntax has impact on the meaning of query terms. Perhaps such research may help determine what possible categories a word can belong to. Perhaps, it also represents a limit to what kind of information the computer can use to assist a user in finding relevant information. If the above analysis is correct, then an IR system may be able to go beyond simply relying on a dictionary to determine the meanings of words. If the ability of a search engine to respond to queries is going to depend on its ability to discern the meanings of words in an

environment where old meanings are vague and new meanings are generated, then we will have to address the issue of determining meaning for individual words.

The above analysis and discussion is an attempt to discover the rich variety of strategies that humans use to induce search engines to cooperate with the human's information needs. The enormous variety of lexical and syntactic patterns employed reflect a confusion on the part of the user on how to best explain the information need to the computer. Some strategies seem to reflect that the user thinks of the search engine as one would a small child who only understands single words and cannot handle the additional information conveyed in complex expressions. Other strategies seem to reflect that the user thinks of the computer as an 'all-knowing' entity that can easily comprehend complex expressions, sorting through the syntax and semantics in much the same way another human would (although faster and with much better access to information). Hopefully, with further syntactic and semantic analysis, we can bridge the gap between user and computer.

REFERENCES

Brajnik, G., Guida, G., & Tasso, C. (1987). User Modeling in Intelligent Information Retrieval. *Information Processing and Management 23*, 305-320.

Croft, W. B, Cook, R., & Wilder, D. (1995). Providing Government Information on the Internet: Experiences with THOMAS. *Proceedings of Digital Libraries '95 Conference* (pp. 19-24).

Jansen, B. J., Spink, A., & Saracevic, T. (1998). Failure Analysis in Query Construction: Data and Analysis from a Large Sample of Web Queries. *Proceedings of the Third ACM Conference on Digital Libraries* (pp. 289-290).

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Searchers, the Subjects They Search, and Sufficiency**:** A Study of a Large Sample of Excite Searches. *Proceedings of WebNet 98 Conference*.

Jones, S., Cunningham, S. J., & McNab, R.(1998). Usage Analysis of a Digital Library. *Proceedings of the Third ACM Conference on Digital Libraries* (pp. 293-294).

Lawrence, S. , & Giles, C.L. (1998). Searching the World Wide Web. *Science, 280*(5360), 98-100.

Lynch, C. (1997). Searching the Internet. *Scientific American, 276*, 50-56.

Moravscik, J. (1990). *Thought and Language.*New York: Routledge.

Peters, T. A. (1993). The History and Development of Transaction Log Analysis. *Library Hi Tech. 42, 11*:2 , 41-66.

Saracevic, T., Spink, A., & Wu, M. M. (1997). Users and Intermediaries in Information Retrieval: What are they talking about? *Proceedings of the Sixth International Conference on User Modeling* (pp. 43 – 51).