AMY WARNER
School of Information and Library Studies
University of Michigan—Ann Arbor

Natural Language Processing:
Current Status for Libraries

ABSTRACT
A general introduction to natural language processing is provided, including a definition and an overview of how natural language processing systems work. Representative systems from both the research and applied sectors are presented in order to illustrate the state of the art in the field and the issues which underlie system design and implementation. Actual and potential areas for natural language processing in information retrieval, including retrieval from online catalogs, indexes, and full texts are discussed, with an assessment of short- and long-range agendas and possible limitations.

INTRODUCTION
Applications of artificial intelligence (AI) to library and information science have been investigated since the late 1970s, and have focused for the most part on expert systems as the most relevant area of AI to pursue. The other papers in these proceedings reflect this interest in expert systems research and development, in their coverage of applications areas (including reference, cataloging and indexing, document delivery, and the user interface); theoretical models (user models); and technologies (knowledge representation techniques). This is understandable given that many of the identified tasks exhibit at least some of the characteristics which make them amenable for expert systems development (Brooks, 1987).
At the same time, however, most of the data manipulated in our automated systems are textual. The information systems themselves consist primarily of free-form natural language text, and they are queried using textual representations as well. Given the sheer quantity of text now available to be searched in machine-readable form, it is not surprising that the "management" of that text by both the system designer and the user is becoming an increasingly difficult problem. This seems most apparent with full text of documents, which are particularly difficult to search and browse given current retrieval techniques (Blair & Maron, 1985).

The main assumption of this paper is that one major problem in human interaction with textual databases is linguistic. Thus, whereas it is very important to understand and model the expert system heuristics associated with the retrieval process, it is also crucial to understand, represent, and effectively manipulate the relevant linguistic structures. This is the problem domain for natural language processing within information retrieval, including interactions with both commercial IR systems and online catalogs.

In addition to justifying this basic assumption, this paper also addresses the following themes and issues:

1. The scope of natural language processing and its relationship to artificial intelligence, specifically to expert systems.
2. The basic architecture of natural language processing systems, and some guiding assumptions, both practical and theoretical, of the field.
3. How and where natural language processing can be most usefully applied in information retrieval.
4. The potential and the limits of natural language processing.

NATURAL LANGUAGE PROCESSING (NLP)

The area known as "natural language processing" is one of three fields which are highly related in their merger of certain aspects of linguistics and computer science. The cognate fields which will be defined include computational linguistics, natural language processing, and natural language understanding. Although these terms often mean somewhat different things to different people and are in fact sometimes used interchangeably, an attempt is made here to make valid distinctions among them by discussing their similarities and differences.

Probably the oldest of these fields is computational linguistics, which is essentially concerned with the algorithms or formalisms that are used to process language, specifically with their computational
power. A major issue involves research into the computational tractability of various linguistic formalisms, a major concern in systems implementation.

Natural language processing is an area of research and application that explores the computer processing of natural language as part of a system that is intended to interact in some way with a user. Input and output may be in the form of single sentences or sentence fragments, or in connected text. Furthermore, language can be entered and retrieved in spoken or written (keyed) form, with this discussion emphasizing the written form.

Natural language understanding is the part of natural language processing which aims at discovering and using knowledge representation techniques from artificial intelligence in language processing systems. These representations are either intended to aid in more flexible, in-depth processing of the linguistic data (an engineering approach) or are intended to serve as psychological models of human language production and comprehension (a cognitive approach) (Hayes, 1978).

Systems which computationally process and manipulate natural language may therefore fall within or outside the AI paradigm depending on whether their algorithms are claimed to "understand" the language being manipulated, in which case they are more accurately referred to as natural language understanding systems. "Understanding" is usually accomplished by using well-known AI data structures such as semantic networks, scripts, and frames.

The term natural language processing is used generally to refer to all technologies and systems, both AI and non-AI based, which analyze and manipulate the linguistic data. All natural language processing systems, whether or not they incorporate AI technology, are built to accomplish some linguistic task, such as text understanding, text generation, and natural language interfaces to database management systems or expert systems (Warner, 1987). The term natural language processing is used most frequently in this paper, in which the main point being investigated is the use of the broad range of pure natural language processing techniques in information retrieval systems.

All three areas—computational linguistics, natural language processing, and natural language understanding—have drawn at various times and to varying degrees on work from linguistic theory. This is the academic discipline which studies and attempts to formally model the structure of language. The computational power and tractability of these formalisms have been investigated by computational linguists, and developers of natural language systems have sometimes based their processing algorithms on them. However, since there is no language whose structure has been completely formalized by theorists, these models often must be greatly extended or modified by natural language
systems developers. There are also systems which are not based on any model from linguistic theory, but whose processing algorithms incorporate new approaches developed from scratch by the AI community.

Another useful distinction to make is between natural language processing and expert systems, since both are often considered components of artificial intelligence and therefore share many things in common. The most important thing they share—that which makes them part of artificial intelligence—is their focus on automating tasks which are believed to require human intelligence. Beyond that, there are some fundamental differences between them which serve to distinguish them as separate enterprises (Mishkoff, 1985):

1. The most overt difference between the two is in their applications. Natural language processing is used to produce natural language interfaces to databases and to process and manipulate the linguistic structures in a text. Expert systems are used to perform the reasoning processes associated with particular technical domains.

2. The domain of natural language processing is human language, whereas the domain of expert systems is some specialized area of human expertise.

3. Language is acquired through a largely unconscious process starting in early childhood, whereas the more specialized knowledge associated with expert systems is acquired later through a conscious learning process. Therefore, the process of discovering the rules to be automated in the two domains is different. The rules of language are indirectly inferred by analyzing linguistic data, whereas the expert system rules are consciously identified by interacting directly with a domain expert in a process known as knowledge engineering.

The preceding discussion implies that these are totally separate areas when in fact they are not. For example, a major effort is underway to endow expert system interfaces with more flexible linguistic capabilities, indicating a merger of the two into one architecturally complete information system (Finin et al., 1986).

Natural Language Processing Systems

Architecture and Issues

The long-range goal of research and development in natural language processing is to endow a computer with all the necessary rules to process language completely. Underlying this goal are the assumptions that language is systematic and rule-governed; that these rules are discoverable through linguistic analysis; and finally, that the rules, once discovered,
are amenable to computational implementation. However, it is clear that
the rules of language are both numerous and complex, and the goal
of a fully flexible natural language system therefore remains a long-
range one. In the short term, parts of the whole problem are being tackled
separately in smaller, more manageable systems.

**Table 1**

**Simplified View of a Natural Language System**

<table>
<thead>
<tr>
<th>Natural Language System Component</th>
<th>Linguistic Level</th>
<th>Applied IR Analogy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological analyzer</td>
<td>Morphological level</td>
<td>Truncation</td>
</tr>
<tr>
<td>Lexicon</td>
<td>Lexical level</td>
<td>Stoplist</td>
</tr>
<tr>
<td>Parser</td>
<td>Syntactic level</td>
<td>Noun phrases</td>
</tr>
<tr>
<td>Semantic analyzer</td>
<td>Semantic level</td>
<td>Thesaurus</td>
</tr>
<tr>
<td>Pragmatic analyzer</td>
<td>Pragmatic level</td>
<td>Thesaurus hedges</td>
</tr>
</tbody>
</table>

The first two columns in Table 1 present a highly simplified view
of the components of a natural language system and the linguistic levels
to which they map; this breakdown is based on the architecture explicated
by Winograd (1984). Five levels are described, exemplified by the
following description of processing of the simple sentence “The system
retrieved relevant articles.”

1. **Morphological.** Words (roughly letters bounded by spaces) are
decomposed into roots and endings. For example, the term “articles”
would be broken into the root “article” and the plural ending “-s.” This is accomplished by the morphological analyzer.

2. **Lexical.** Using a dictionary, each root is assigned a set of lexical
categories. For example, the stem “article” would be assigned the
lexical category NOUN through look-up in a lexicon.

3. **Syntactic.** Using a program module called a parser, a structural (i.e.,
grammatical) description is assigned to the sentence. The program
takes the word level input of the lexical component and decides how
the individual words go together to form phrases, clauses, and whole
sentences. The following analysis of the sentence would result:

```
S
  NP
    ART THE
    N SYSTEM
  VP
    V RETRIEV-ED
    ADV RELEVANT
    N ARTICLE-S
```
4. **Semantic.** The syntactic structure is translated into a form which represents the meaning of the sentence. This involves the determination of the appropriate meaning of each word and then the combination of these into some logical form. This will allow certain inferences to be made about the input. For example, given an appropriate semantic representation, the sentence "The system retrieved relevant documents." would be interpreted as true, since the system would "know" that articles are kinds of documents by interacting with some knowledge representation scheme, such as a semantic network which stores that information.

5. **Pragmatic.** This analyzes the sentence in its context, taking into account a certain body of knowledge about the domain and about the plans and goals of the speaker (user) and hearer (computer) in the conversation. For example, pragmatic information would allow the system to infer that a computer was involved in the retrieval operation, although it is not explicitly stated in the sentence.

The architecture just described should not be considered standard. It is based on the notion that there is so much going on in language that it is necessary to focus on one level of the structure at a time (Crystal, 1987, pp. 82-83). The conception is that a natural language processing system should be modularized—that is, the sentence is processed entirely at one level before being passed on to the next level. This is an intuitively appealing approach, since it allows the designer to work on each smaller component of the system in isolation, and, conversely, to locate and correct errors more easily. However, another more recent approach is described by Allen (1987), in which partial results are passed between modules before analyzing the entire sentence. Furthermore, not all systems contain all the modules delineated above. For example, some combine the syntactic and semantic analysis to produce a semantics-driven parser, as in the system described by Schank and Birnbaum (1984). Finally, systems do not all process to the same depth in terms of linguistic levels, with morphology being the shallowest and pragmatics being the deepest (Weischedel, 1986).

The range of capabilities of current systems is described by Warner (1987), who also provides a summary of issues which have guided research and development in the area. These issues include the following:

1. **Robustness.** Research and development in natural language processing has been oriented toward producing systems with greater depth of analysis and flexibility. Work in this area focuses on processing of ungrammatical or partial input (Carbonell & Hayes, 1984); novel language including metaphor (Carbonell, 1982); and the context of sentences or texts, including the goals and plans of the

2. **Transportability.** Since natural language processing systems can now operate only in limited subject domains, one of the greatest problems is how to best transport techniques used in one subject domain to a new one. This involves not only a system design which makes it transportable, but also a method for customizing it to the new environment (Marsh & Friedman, 1985; Grosz, 1983).

3. **Sublanguage analysis.** At present, some natural language processing systems are being built to process text in small subject domains (e.g., medicine, molecular biology, etc.) characterized by a subset of linguistic patterns—i.e., characteristic constructions. This effectively reduces the number of operations which must be coded and carried out to a manageable size. One long-standing project based on sublanguage analysis is New York University’s Linguistic String Project (LSP) (Sager, 1981; Sager et al., 1987). It uses a precise sublanguage description to convert hospital records into a structured format, which can then be used in various applications, including the production of summary reports and question answering.

4. **Ambiguity and synonymy.** A major theme in natural language processing centers around the processing of specific constructions which are known to be either highly ambiguous or synonymous with other constructions. The goal is to endow the system with the capability to generate only one analysis for each linguistic structure (resolve ambiguity), and to generate the *same* analysis for different structures which have the same meaning (eliminate synonymy). This effectively results in a one-to-one correspondence between form and meaning. Constructions in which ambiguity or synonymy need to be handled include compound noun phrases (e.g., FOREIGN STUDENT TEACHING—is this "teaching of foreign students" or "teaching by foreign students"?) (Sparck Jones, 1985; Taylor et al., 1989); coordinate constructions (e.g., OLD MEN AND WOMEN WITH GLASSES—are both men and women old or is it just the men who are old?) (Fong and Berwick, 1985); and paraphrases (JOHN HIT THE BALL/ THE BALL WAS HIT BY JOHN—roughly synonymous structures analyzed the same way: JOHN (agent) HIT (verb) BALL (patient) (Harris, 1985, pp. 326-29).

In summary, there is a need within natural language processing to build flexible, cooperative systems based on rules which can be used in other new systems. However, because language is so complex and ambiguous, this can only be done currently in limited domains and only for certain constructions. An important generalization underlies this: There is, in general, a trade-off between the subject breadth of
the information contained in the system and the depth of processing which can be performed on that information. Essentially, very deep (i.e., pragmatic) analyses can only be carried out in highly restricted subject domains, while greater subject breadth means that the analysis will be shallower.

Operational Information Retrieval System Parallels

Many of the natural language processing systems just covered could be considered information retrieval systems. Indeed, one of the applications areas within that field is the design of natural language interfaces. However, these serve as front ends to database management systems and expert systems rather than to document retrieval systems. This section surveys the parallels between pure natural language processing and document retrieval from bibliographic databases (online catalogs and indexes) and full-text databases.

It is useful to begin by summarizing the current capabilities and architecture of applied (i.e., nonexperimental) document retrieval systems and interfaces to these systems. This discussion, which is an elaboration of the material in column one of Table 1, is based largely on the overviews presented by Doszkocs (1986, 1987) and also refers to the examples in Table 2.

1. Morphological level. IR system capabilities for dealing with morphology include prefix, infix, and suffix truncation operators. This means that the system will “ignore” the affix in its matching process. The examples in Table 2a illustrate this.

2. Lexical level. IR systems do not use a lexicon to assign parts of speech to individual stems, as in the natural language processing systems previously described. However, a stopword list is employed to prevent machine indexing of certain function words which are not considered useful for content representation.

3. Syntactic level. Structural units above the level of individual words or stems are primarily noun phrases. Noun phrases are found at two places in the retrieval system operation. They are part of the system’s inverted indexes, but only if they come from controlled term fields (i.e., descriptors, subject headings, or identifiers). They are also “constructed” at search time through insertion by the user of proximity operators which will allow noun phrase variants to be retrieved. For example, the search term PROGRAMMING (2N) INTERFACE would retrieve documents containing any of the syntactic paraphrases found in Table 2b.
4. **Semantic level.** IR systems allow users to manipulate meaning relationships among terms through certain interactions with the thesaurus. For example, given the equivalence relationship found in Table 2c.1, some retrieval systems will automatically substitute the preferred term in the user's strategy if necessary. Also, given the BT-NT relationship depicted in Table 2c.2, in some systems which contain an online thesaurus users can expand their requests by automatically incorporating terms from the hierarchy.

5. **Pragmatic level.** In some ways, the thesaurus can be said to contain pragmatic information. This is because many of the decisions about the relationships among terms are based on indexing practice. For example, the instruction to index a surgical procedure with an accompanying body part (Table 2d) really pertains to the pragmatic level. Another example of pragmatic information in retrieval systems might be the "hedges," collections of search terms associated with particular topics, which have been found through experience to successfully retrieve relevant documents (Sievert & Boyce, 1983).

| Table 2  
| IR Data from Linguistic Levels |
|---|---|
| a. Morphology | Graphic interface/Graphic[al] interface  
Nonlinear operat[or]/Nonlinear operat[ion] |
| b. Syntax | Programming interface  
Interface for programming  
Interface for computer programming |
| c. Semantics | 1. Syntax/Grammar  
2. Computer interface  
IBM interface  
Macintosh interface |
| d. Pragmatics | Cataract extraction + Lens, crystalline |

Operational information retrieval systems can be discussed in terms of the issues of robustness; domain breadth and processing depth; transportability; and handling of ambiguity and synonymy. Most of them operate in very wide subject domains in which large amounts of textual material are processed and in which a wide variety of linguistic constructions are potentially available to be manipulated. However, processing is quite shallow and usually involves the isolation and storage of individual words (strings of characters bounded by spaces); phrases are only isolated and stored when they have been previously assigned from a controlled vocabulary by an indexer or cataloger. Very few incorporate any semantics, although the thesaurus, if available online, is relevant only to a particular database and contains some information about semantic relationships among terms, such as synonymy and genus-species; however, the thesaurus is frequently not linked to the database.
and the user must often explicitly select terms from the displays. At the same time, most commercially available "user-friendly" interfaces to online databases have exploited the simple internal structure of the database and similarly employ very shallow linguistic analysis. A typical interface might allow the user to input a string of search terms, which would then be searched by automatically inserting a Boolean or single adjacency operator (Benson & Weinberg, 1988) or by automatically stemming and weighting the individual words in the query (Koll et al., 1984). Only a few, such as Tome-Searcher ("Intelligent Search Software...", 1988), provide for query expansion using a lexicon. However, none provides the range of pragmatic query broadening and narrowing capabilities detailed in the exploratory study of Fidel (1986).

Although commercial systems and interfaces are not very robust (i.e., they do not process very deeply or flexibly), their algorithms are very domain-independent and therefore transportable. This is because they have, except in the cases where vocabularies are linked to the databases, worked by simple surface matching of character strings in queries and documents—such an algorithm makes use of no deep "knowledge" of the linguistic or contextual knowledge of the particular domain.

Ambiguity and synonymy have been major issues in information retrieval, and their resolution is one of the major functions of a controlled vocabulary (Lancaster, 1979, p. 181). Ambiguity and synonymy of natural language search terms in isolation are usually resolved when combined with other terms in the query, and are therefore not considered too problematic in the operational information retrieval environment.

NATURAL LANGUAGE PROCESSING AND EXPERIMENTAL INFORMATION RETRIEVAL

Although much of the current interest in producing more sophisticated IR systems has focused on expert systems development, there has been a historic connection between natural language processing and information retrieval. This was investigated by Masterman, Needham, and Sparck Jones (1959), who stated that:

An analogy made between library retrieval and mechanical translation is usually made by assimilating library retrieval to mechanical translation. We desire to draw the converse analogy; that is to assimilate mechanical translation to library retrieval. To do this, mechanical translation procedures must be generalized and made interlingual, until they become as general as library retrieval procedures already are. This generalization can be made if the mechanical translation procedure is based on a thesaurus. (p. 917)

This idea reflected the thrust of efforts in machine translation, which was based on look-up of individual words in dictionaries. Although
this was very appealing, it was also simplistic and therefore not very successful.

Natural language processing research continued, but instead focused on isolating and manipulating complex linguistic structures for other applications, such as question answering, rather than matching and translating individual words. The work in the area was spawned by the formalism in linguistic theory known as transformational-generative grammar (Chomsky, 1957), which seemed amenable to computational implementation. This sparked a number of attempts within information retrieval to directly automate formalisms from linguistic theory in order to improve system performance in areas such as automatic indexing and automatic abstracting. Surveys exploring the relevance of linguistic theory and information retrieval were conducted by Sparck Jones and Kay (1973, 1977) and Montgomery (1972); actual experimental systems based on linguistic theory were implemented by investigators such as Moyne (1968). Since the 1970s, there has been a trend away from direct implementation of formalisms from linguistic theory in IR systems, and a trend toward the adoption of AI approaches (Sparck Jones & Tait, 1984; Croft & Lewis, 1987) as well as the empirical discovery and development of non-AI algorithms tailored to a given retrieval problem or environment (Salton, 1989; Dillon & Gray, 1983; Schwarz, 1988).

Two important points which are relevant to the role of natural language processing to information retrieval need to be made. First, searches of retrieval systems are usually by some topic and are intended to retrieve a set of documents which match a request; this is in contrast with much of the work in pure natural language processing, where systems are often intended to answer specific questions by retrieving particular facts from a database. Second, it is generally assumed, at least within applied IR, that the subjects of documents and queries can be represented adequately by lists of words and phrases; this contrasts with other natural language processing systems in which the linguistic information in the system often results from the full processing of linguistic units at the sentence level and above (i.e., connected text). These fundamental differences have prompted some (Lancaster, 1972, p. 141; Salton & McGill, 1983, p. 258) to question whether natural language processing is relevant to document retrieval. However, these statements were made in an era characterized primarily by intermediary searching of bibliographic databases.

As useful as most current operational retrieval systems and interfaces are, more recent developments in interfaces for end-users and full-text retrieval have revealed a need for even more powerful retrieval aids. Efforts to produce them have been the major focus of experimental
information retrieval, and some investigators are making use of natural language processing techniques in those endeavors. There are four general areas which are of current concern:

- Making the systems “transparent” (Williams, 1986), in which more functions would be delegated to the machine, has become a primary goal of the effort to design powerful interfaces.
- Systems with interface capabilities also could be further enhanced by more robust processing (i.e., phrase as well as keyword indexing) of the underlying free text in titles and abstracts.
- Since retrieval effectiveness does not appear to “scale up” very well to large full-text databases (Blair & Maron, 1985), another major issue has involved the manipulation of these texts into a representation which can be searched more effectively.
- The costs of manual production and application of controlled vocabularies, which have always been high, could be lessened through effective automatic procedures.

Comparison of Experimental and Operational Systems

It is useful to compare experimental and operational information retrieval in terms of the analysis procedures which are employed at the five linguistic levels already described. This results in a view of the state of the art of experimental information retrieval and a vantage point from which to discuss both future directions and limitations.

Morphological and lexical analysis within experimental information retrieval closely parallels the procedures employed within operational systems. Stoplists of terms are used in experimental systems to exclude frequently occurring, primarily function words which are not considered to be indicative of subject content. In contrast with operational systems where users normally have to supply truncation operators, automatic term truncation is virtually standard in experimental systems. Virtually all systems which perform automatic indexing and/or natural language query manipulation make use of a truncation procedure at an early stage in their algorithms. It is important to note that automatic truncation procedures in experimental document retrieval systems do not usually employ a fully developed morphological analyzer, as is the case in most of the natural language processing systems previously described. Instead, they linguistically overgeneralize, using, for example, a list of suffixes to remove the longest matching suffix on the end of a given word; this results in an efficient processing algorithm, although it does produce some processing errors (Salton & McGill, 1983, pp. 72-73).

Syntactic analysis within experimental information retrieval in general focuses on the isolation of noun phrases from free text—titles,
abstracts, full texts, and natural language queries. This is achieved through some kind of parsing procedure, although it is not necessary to process any given sentence as fully as in other natural language processing systems. One purpose of a syntactic analysis of this sort is to enable a system to process strings of natural language words input by the user into meaningful phrases; this can be used as a precision device (Metzler & Haas, 1989) or as a method of grouping related phrases for query reformulation (Salton, 1989). Another purpose is to automatically index a document collection using noun phrases instead of the usual keywords (Dillon & Gray, 1983; Schwarz, in press).

Work in incorporating semantics into experimental information retrieval systems has been undertaken for a variety of purposes and in a number of ways. Semantic analysis is useful in retrieval systems since it permits word and phrase manipulation based on criteria other than the matching of surface strings; that is, it is an attempt to manipulate word senses rather than word tokens. Thus, in order for semantic analysis to be accomplished, information regarding the meanings of terms and/or their relationships to each other must be identified, stored, and made available to the system. In one approach, investigators build semantic representations to be used to manipulate queries in interface design; semantic representations built for this purpose are quite varied, and include case frames, fuzzy logic, and semantic word classes (Croft & Lewis, 1987; Biswas et al., 1987; Liddy & Jorgensen, 1989). Systems which attempt to semantically process both the query and the document store are also based on different techniques, ranging from semantic analyzers which manipulate semantic primitives (Sparck Jones & Tait, 1984) to parsing procedures which extensively employ semantic information from available machine-readable dictionaries and thesauri (McCray, 1989). Finally, attempts are being made to use semantic criteria to automatically construct a thesaurus for an information retrieval system from a machine-readable dictionary (Ahlswede et al., 1988).

The pragmatic level has been explored informally by a few investigators. The discourse properties of scientific abstracts have begun to be explored (Liddy, 1988; Liddy et al., 1987), specifically to determine if there is any regular, implicit structure which can be exploited and whether there is any predictable way to determine the anaphoric referents (e.g., the specific entity to which a pronoun refers). The determination of such regular structures would be useful in both searching and automatic indexing procedures. Other projects are experimental systems which deal with natural language input but also consider search strategy formulation and reformulation an expert system task. For example, the
IR-NLI-II system (Brajnik et al., 1987) incorporates mechanisms which allow for the understanding of anaphoric referents and indirect speech as well as the management of a clarification dialogue with the user and the IR system (Croft & Thompson, 1987) creates a model of the user's information need which can be modified based on evaluation of system output and/or a change in the user's goals.

Virtually all of the issues previously identified for pure natural language processing also apply to its application within experimental information retrieval. Automatic linguistic techniques which are generalizable across wide subject domains primarily exist at the lower levels (e.g., morphology and syntax). Systems employing semantic techniques often contain domain-dependent information which usually needs to be hand constructed, making them much smaller and therefore much less extensible to an operational environment. An improvement over this situation is the more recent trend toward exploiting machine-readable dictionary information (Krovetz & Croft, 1989), which enables the system to have at its disposal much more semantic information. Furthermore, systems developers are concerned with many of the difficulties associated with constructions which are either ambiguous or synonymous and therefore difficult to process effectively. These include ambiguous noun phrases (McCray et al., 1988); noun phrases and their paraphrases (Dillon & McDonald, 1983; Salton, 1989; Sparck Jones & Tait, 1984); and conjunctions (Das-Gupta, 1987).

In summary, the following generalizations may be made about natural language processing and document retrieval, including both experimental and operational systems: first, there continues to be a reliance on subject representation techniques by words and/or phrases. Operational and experimental systems differ not in what they represent, then, but in terms of the degree to which they isolate, manipulate, and interrelate these structures. Thus, experimental systems tend to be much "richer" in their processing of syntactic units (noun phrases) and semantic (word and phrase) senses and their relationships. Second, in general, size and domain breadth remain problems within information retrieval because they create an unfortunate trade-off with processing depth. This means that domain independent analysis, generalizable across large numbers of operational systems, remains largely dependent on the matching of surface strings, with surface analyses and matches often not fully accounting for synonymy and ambiguity in lexical items, which often can be resolved only at a deeper, semantic level (e.g., GRAMMAR and SYNTAX from Table 2c.1 cannot be given an equivalence relationship automatically through a simple match of any surface elements).
AN AGENDA FOR NLP AND INFORMATION RETRIEVAL

The history of information retrieval has demonstrated that the role for natural language processing within the field is a controversial one. It is still fair to say that natural language processing techniques remain largely a promissory note rather than an accepted and established agenda. Although this can be considered a harsh indictment, it does reflect the larger problem of how to meaningfully relate the procedures and findings of the research domain with what is going on in the applied arena. At the same time, however, it is clear that system performance is still far from perfect (i.e., 100 percent recall and precision) and that other, nonlinguistic techniques have not improved retrieval performance very much (Lewis et al., 1989).

In many ways, one can view the linguistic goal in all information retrieval endeavors as the elimination of linguistic variability—that is, eliminate ambiguity and synonymy of search and index terms and create a one <--> one relationship between term forms and the concepts they represent. In the realm of operational information retrieval, construction of controlled vocabularies and their application can be seen as an attempt to manually reduce this variability. Document representation and searching by uncontrolled natural language have other, complementary benefits over indexing and searching with controlled vocabularies (no difficulties associated with the imposition of an artificial language), but they have some disadvantages as well, particularly that users are given the task of handling the linguistic variability of the underlying text themselves by, for example, supplying all synonyms for a given concept. Lancaster (1979, pp. 284-88) noted the trade-offs between natural language and controlled vocabulary, and recognized the need to control natural language more effectively at search time.

Research in experimental information retrieval can be seen as an attempt to eliminate the same kinds of variability which operational systems have eliminated by using indexing languages (controlled vocabulary) or by relying on the searcher to cope with the variability (natural language). Techniques borrowed from linguistics and natural language processing still seem to offer great promise for discovering and automatically managing and manipulating the variability of the natural language of texts and queries.

A continuing problem is the barrier imposed by the domain breadth/processing depth trade-off. This means that the easiest and most computationally viable linguistic techniques to use within information retrieval remain those from morphology and syntax. It is possible that Lewis, Croft, and Bhandartru (1989) are right—that surface (morphological and syntactic) techniques unequivocally will not result in very great improvements in retrieval performance, and that they should be
abandoned in favor of semantics. However, this seems to imply that the relevant linguistic problems in retrieval lie either on the surface (grammatically) or deep (semantically) inside the text. In fact, there do seem to be some phenomena which do lie on the surface, as the examples in Table 2a, b illustrate. Thus, it seems reasonable to suggest that grammar does play a role in some of the linguistic variability which should be accounted for in the retrieval environment; this in fact has been the viewpoint of some within the natural language processing community for quite some time (Marcus, 1984).

Nevertheless, that semantics accounts for much of the language variability in information retrieval has been well documented by Blair and Maron (1985) and Sievert and McKinin (1989) for full text, and by Lesk (1988) for catalogs. Basically, the problem is that, given a particular concept in retrieval, the phrasing of that concept is highly variable; this is not just a syntactic problem, as already pointed out, but can also be a semantic one as the example in Table 2c.1 illustrates. This justifies the techniques which store and manipulate the senses of words and phrases. However, automatic semantic techniques require that elements of meaning and their relationships be accessible from some source. The necessity of deriving this from scratch has made these systems small, and they will probably remain small unless some way is found to produce larger semantic information stores. There are basically two ways of doing this:

1. Extract the semantic information automatically from a machine-readable dictionary. There are several machine-readable dictionaries now available, and a current topic of intense investigation within the natural language processing literature involves the automatic extraction and use of information from these tools (Byrd et al., 1987; Boguraev & Briscoe, 1987).

2. Employ large-scale manual procedures to explicitly encode certain linguistic features of texts and/or terminology which can then be manipulated automatically. For example, the Unified Medical Language System (UMLS) (Tuttle et al., 1988) is a project sponsored by the National Library of Medicine, which is attempting to build a "meta-thesaurus" of biomedical terminology which can be used to provide a uniform user interface to heterogeneous sources of information.

CONCLUSION

The preceding discussion presented what has legitimately been described as a simplistic view of the retrieval situation. It implies that the user and the system (i.e., the speaker and the hearer) understand
concepts and texts in the same ways. Among others, Belkin, Oddy, and Brooks (1982) have advocated the design of retrieval systems which can build models of the user’s needs and views of the world and can revise that model based on additional information. Some individuals would call this an expert retrieval system. However, it is clear that this expert system would need large amounts of pragmatic information about what constitutes a cooperative dialogue and about how to diagnose and correct retrieval errors based on notions such as what the user wants from the system and what the user knows about the system.

Thus, an agenda for the development of increasingly sophisticated information retrieval systems which incorporate natural language processing techniques can be proposed. It involves investigation of linguistic phenomena as well as implementation and testing in actual systems of structures from all linguistic levels: morphological, syntactic, semantic, and pragmatic. Furthermore, it seems reasonable, at least in the short term, to implement a strategy in which as much is done on the surface (with morphology and syntax) as possible, leaving semantic and pragmatic analysis for problems which cannot be solved in any other way.

The prospects for making progress in all of these areas now seem better than ever because of the strides which have been made in natural language processing itself, which is now a more mature field; the increasing amount of study within linguistic theory of semantic and pragmatic phenomena (Morgan, 1982) and the structure of texts (Beaugrande & Dressler, 1981); and the favorable climate within information retrieval for a new look at linguistic techniques and what they have to offer (Croft, 1987).

**Future Prospects**

An open question always remains about how far information retrieval can go with linguistic techniques. The goal of a fully automatic, fully flexible retrieval system may never be realized; however, systems can surely be made more flexible, adaptable, and responsive than they are, and we can also learn something about the linguistic structures inherent in texts and queries in the process.

In the end, however, information retrieval, having a large applied component, will judge any product by its utility and not by whether or not it is a system based on expert systems, natural language processing, or any other technology.

As with any computer product, the value to the user has nothing to do with the underlying techniques used to create the product. The user just wants something to solve a problem, and a product either solves it or doesn’t. If it does answer a need, the product must be judged in its effectiveness against other products that solve the same need. (Harris & Davis, 1986, pp. 156-57)
REFERENCES


from a medical to a Navy domain. *ACM Transactions on Office Information Systems*, 3(2), 121-140.


