The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives

Jane Greenberg

Abstract

Natural language processing (NLP) is an extremely powerful operation—one that takes advantage of electronic text and the computer's computational capabilities, which surpass human speed and consistency. How does NLP affect archival operations in the electronic environment? This article introduces archivists to NLP with a presentation of the *NLP Continuum* and a description of the *Archives Axiom,* which is supported by an analysis of archival properties and objectives. An overview of the basic information retrieval (IR) framework is provided and NLP's application to the electronic archival environment is discussed. The analysis concludes that while NLP offers advantages for indexing and accessing electronic archives, its incapacity to understand records and recordkeeping systems results in serious limitations for archival operations.

Introduction

Since the advent of computers, society has been progressing—at first cautiously, but now exponentially—from a dependency on printed text, to functioning almost exclusively with electronic text and other electronic media in an array of environments.[1] Underlying this transition is the growing accumulation of massive electronic archives, which are prohibitively expensive to index by traditional manual procedures. Natural language processing (NLP) offers an option for indexing and accessing these large quan-

[1] The electronic society's evolution includes optical character recognition (OCR) and other processes for transforming printed text to electronic form. Also included are processes, such as imaging and sound recording, for transforming and recording nonprint electronic manifestations, but this article focuses on electronic text.

The author would like to thank Professors Richard Cox and Edie Rasmussen of the School of Information Sciences, University of Pittsburgh, for their encouragement with this topic; and the external reviewers for their helpful suggestions.

tities of electronic archives. While NLP has been researched and used in various domains since the 1950s,[2] its application in the electronic archival frontier is more recent[3] and invites a host of questions. What are the strengths and weakness of applying NLP in the electronic archival environment? Should archivists be spending their limited resources and energy investing in NLP? What steps might archivists take to develop electronic recordkeeping systems that intelligently employ NLP? Archivists need to explore these and other related questions.

This investigation introduces archivists to NLP with a presentation of the *NLP Continuum* and a description of the *Archives Axiom*, which is supported by an analysis of archival properties and objectives. An overview of the basic information retrieval (IR) framework is provided and NLP's application to the electronic archival environment is discussed. The appendix to this article contains a bibliography for archivists who want to further explore NLP.

What Is NLP?

Tamas E. Doszkocs provides the following definition:

> Loosely defined, natural language processing (NLP) encompasses all computer-based approaches to handling unrestricted written or spoken language, from purely "mechanistic" procedures, as employed by text editors, word processors, and in automatic-indexing approaches in information retrieval (IR) to "intelligent" analysis, understanding and expression of "meaning" as exemplified in natural language understanding, question answering and expert systems (AI).[4]

Doszkocs's definition indicates the broad spectrum of information processing activities to which the abbreviation NLP is freely applied. Essentially, his definition encompasses the *NLP Continuum*[5] (Figure 1).

The four key points along the *NLP Continuum* are not mutually exclusive; rather, NLP activities can take place anywhere, combining various character-

[2] "Computers, as automatic symbol–manipulation machines, have been used for the processing of natural language virtually from their earliest introduction in the 1950s." Tamas E. Doszkocs, "Natural Language Processing in Information Retrieval," *Journal of the American Society for Information Science* 37 (July 1986): 191.

[3] One example is the HELIOS project, which involves the archives of Senator Henry John Heinz III (R-PA). The archives have been scanned via OCR and are accessible by CLARIT, a sophisticated NLP software developed by the CLARITECH Corporation. See Edward A. Galloway and Gabrielle V. Michalek, "The Heinz Electronic Library Interactive Online System (HELIOS): Building a Digital Archive Using OCR, and Natural Language Processing Technologies," *The Public-Access Computer Systems Review* 6, no. 4 (1995). Available at <http://info.lib.uh.edu/pr/v6/n4/gall6n4.html>.

[4] Doszkocs, "Natural Language Processing in Information Retrieval," 191.

[5] The idea of a continuum evolved from a discussion with Professor Edie Rasmussen, Department of Library and Information Sciences, School of Information Sciences, University of Pittsburgh, 5 March 1995.

		SOPHISTICATED NLP/IR ↓		
Basic	**NLP/IR**	•	**NLP/Rudimentary AI**	**NLP/Full AI**
NLP			**(Syntax, Semantics, etc.)**	**(Discourse)**

[AI= Artificial Intelligence]

FIGURE 1. NLP Continuum

istics. For example, "NLP/IR" can be combined with "NLP/Rudimentary AI" to create what is known as "Sophisticated NLP/IR," a point also marked on the continuum in Figure 1. A brief explanation of the continuum's four key points will help to further define NLP characteristics.

Basic NLP

Basic NLP activities take place when the information seeker's query, represented by a natural language term or a group of terms, freely searches the text of a collection of titles, bibliographic records, abstracts, full text documents, or other document representation forms (hereafter the generic term *document* can refer to any document representation). A basic NLP search may be matched against a document's free text and its controlled vocabulary at the same time, a technique often referred to as keyword searching. This diverse application may help to explain why library, information science, and other related literature often use the abbreviation *NLP* and the terms *natural language*, *free text*, and *keyword* interchangeably. Basic NLP may also be applied to a single instance of a document representation. The execution of a "find" or "find and replace" command during word processing provides an example of basic NLP within a single document.

NLP/IR

NLP/IR involves adding IR processing capabilities to basic NLP to enhance the retrieval operation. Some examples of IR processing capabilities that can be part of an NLP/IR operation include the following:

1. *Boolean, adjacency, and proximity operators*, which refine a search with verbal commands, such as "and," "or," "near," and "within five words of," when two or more search terms are used;
2. *Stemming algorithms*, which truncate suffixes, and/or eliminate prefixes and infixes from search terms;
3. *Stop word algorithms*, which eliminate from the processing activity words without any content value, such as "the," "an," and "a";

4. *Lexical variation enhancements*, which add more relevant terms to the query operation while working with a structured vocabulary tool or a thesaurus similar to *Roget's*;[6]
5. *Term frequency calculations*, which involve statistical algorithms that count the number of times a term or a group of terms appears in a document, usually in relation to the larger document collection or a single document (the result is often referred to as a query-document similarity measure or calculation);
6. *Term weighting algorithms*, which assign numerical values or weights to terms on the basis of their content value (names of key persons, places, and main events, and significant content terms receive higher numerical values than all other document terms, and conjunctions and articles are generally assigned no value at all and are often eliminated from the query-document similarity calculation); and
7. *Clustering algorithms*, which categorize documents or terms on the basis of document attributes (e.g., keywords, authors, citations, or other major document attributes).

Note that the above IR techniques can be, and often are, combined in various combinations for a single NLP/IR operation. Additionally, many of the combined techniques that use a statistical algorithm (e.g., frequency calculations) support a ranked retrieval, ordering results from the most relevant to the least relevant.

NLP/Rudimentary AI

NLP in the rudimentary artificial intelligence (AI) domain deals with formal language analysis and understanding. Machine translation provides an example of NLP at the rudimentary AI level. Features in this environment include processing at the following six levels:

1. *Phonological processing*, which involves interpreting speech sounds in auditory documentation (this level is not specifically addressed in this investigation);
2. *Morphological processing*, which involves parsing word segments such as prefixes, suffixes, infixes, root words, and compound word parts;
3. *Lexical processing*, which involves sensing word meanings (e.g., synonyms, antonyms, or other types of semantic relationships) and can include the creation or use of lexical tools such as a thesaurus;
4. *Syntactic processing*, which involves identifying and parsing grammatical structures in text;

[6] Structured thesauri are constructed according to national and international standards. For example, ANSI/NISO Z39.19–1993, *Guidelines for the Construction, Format, and Management of Monolingual Thesauri* (Bethesda, Md.: NISO Press, 1994). Many editions of *Roget's Thesaurus* have been published. The most up-to-date version is *Roget's Internet Thesauri of English Words and Phrases* (Nothing Limited, webmaster@thesaurus.com, 1997), available at <http://www.thesaurus.com>.

5. *Semantic processing*, which involves interpreting the contextual meaning of words, phrases, sentences, and other grammatical structures in text; and
6. *Pragmatic processing*, which involves creating a knowledge base to facilitate disambiguation.[7]

These six processing levels incorporate many NLP/IR techniques, and they build upon each other and can be combined in ways similar to that observed in the NLP/IR environment. However, NLP in the rudimentary AI domain differs from NLP/IR in that it aims to preserve syntax and semantics to disambiguate the text's message.

The employment of the various NLP/rudimentary AI processing levels is very much dependent on the knowledge domain to which the NLP operations are applied. For example, in a more general domain environment such as a digital collection of employee newsletters or annual reports, an NLP operation will emphasize the syntactic level. This is because the general domain's lack of a distinct boundary invites a plethora of linguistical ambiguities, and identifying semantic rules for knowledge base construction presents an endless, if not impossible, task. However, in a domain-specific environment such as a digital collection of research documents specifically on blood diseases or petroleum investments, NLP operations can emphasize the semantic level: Language ambiguities are limited in this environment, and identifying semantic rules for knowledge-base construction, therefore, presents a more realistic task.

NLP/Full AI

NLP at the full AI level involves discourse or full natural language understanding. Imagine that you are asked, "What did you have for lunch yesterday?" You reply, "I had a ham sandwich." The questioner then responds with, "Did you have Virginia baked ham or Black Forest ham?" You continue to provide more information about your lunch. In this scenario the questioner and you are having a conversation as a result of both the "questioner's" and "your" natural language understanding. Essentially you would both be performing NLP activities to support the conversation. This type of discourse represents the highest level of NLP as it involves *learning* and *intelligence.*

The *notion of discourse* was coined by Alan Turing in 1950. At that time, Turing, a pioneer in communication theory, introduced his Turing test, which stated that "if a program can fool the human into believing another

[7] For a further explanation of these NLP levels see Gerald Salton, *Automatic Text Processing* (Reading, Mass.: Addison-Wesley, 1989), 377–424. Most basic NLP or AI texts covering NLP will provide a discussion of these processing levels.

human is responding then the program is judged as intelligent."[8] Examples of full AI developments (programs supporting discourse) include robots that deliver mail or clean up toxic waste sites; computer programming developments such as IBM's Deep Blue, the computer that defeated the master chess player Gary Kasparov in May 1997; or even telecommunication developments such as the rover Sojourner, which landed on the planet Mars on July 4, 1997, and communicated with the Jet Propulsion Laboratory, California Institute of Technology. (Various AI communities continue to argue as to what constitutes full AI.)

Clearly, the above examples and other full AI developments are not without their limitations, and research satisfying the Turing test has a long way to go. The immense challenge of securing machine intelligence, however, should not keep archivists from exploring NLP operations. Rather, archivists are obligated to explore NLP's applicability for today's exponentially growing electronic archival environment; one of the first critical steps to achieving this end is to reach a consensus with respect to the definition of archives.

What Are Archives?

Archives, for the purpose of this article, are defined in the axiom shown in Figure 2:

Premise 1:	Archives are records.
Premise 2:	Records are evidence of transactions.
Premise 3:	Evidence of transactions is contingent upon the properties of content, structure, and context.[9]
Premise 4:	The properties of content, structure, and context are contingent upon the preservation of institutional* or personal records and their respective recordkeeping systems.

[Archives are records of evidence of transactions that are contingent upon the preservation of content, structure, and context of institutional and personal records and their respective recordkeeping systems.

[*Institutional records may also be identified as corporate or organizational records (e.g., hospital, bank, museum, etc.)]

FIGURE 2. The Archives Axiom

From this axiom it follows that the preservation of a transaction's content-like, structural, and contextual properties is critical to the identification and purpose of archives. In other words, if the properties of content, struc-

[8] Thomas Dean, James Allen, and Yiannis Aloimonos, *Artificial Intelligence Theory and Practice* (Redwood City, Calif.: The Benjamin/Cummings Publishing Company, Inc., 1995), 8. This quote exemplifies one version of the Turing test.

[9] A corollary to premise three could state the following: Evidence is an archival objective; the evidential objective is essential for fulfillment of the accountability and memory objectives.

ture, and context are not preserved, evidence is absent, records are lost, and archives, satisfying the axiom definition presented above, cannot exist.[10]

Archival Properties: Content, Structure, and Context

Archives are defined by properties just as other entities are defined by their unique properties. Consider a chair: The properties of content, structure, and context apply. A chair's content includes the material of which it is composed, such as wood or metal; its structure includes stylistic and physical characteristics, such as ornamentation or height; and its context is represented by how it functions with various entities in its surrounding environment, such as a table or a person who sits on the chair.

With archives, *content* is the symbols (letters, numbers, etc.), words, images, and sounds that constitute the body of the record.[11] Content indicates, to some degree, what a work is about. Consider an official memorandum in an institution's archives (because the topic of this paper is NLP, the memorandum and any other archival records discussed hereafter are assumed to be electronic). The memorandum is composed of words and perhaps numbers and images. These features form the informational content that helps to tell what the memorandum is about. A full and accurate interpretation of content, however, is also dependent upon the memorandum's structural and contextual existence.

Structure explains the stylistic formalism and the physical structure that identifies an archival record.[12] The official memorandum has a very specific style in that it is direct, to the point, and usually brief. The style of a memorandum differs from that of other archives such as the minutes of a meeting, which usually provide an exact account of who said what, and actually may be quite lengthy.

The official memorandum also has certain physical features that allow it to be identified as a memorandum. For example, labels such as "date," "to," "from," and "regarding" are quite common, although different label terminology may be used. These physical features distinguish the memorandum from the minutes of a meeting, which document the "date," "time," "lo-

[10] For a discussion on the concept of a "record" and "recordkeeping systems" see the following works by Richard Cox: "The Record in the Information Age: A Progress Report on Research," *The Records & Retrieval Report* 12 (Jan. 1996): 1–16; "The Record: It is Evolving?" *The Records & Retrieval Report* 10 (March 1994): 1–16; "What is an Archival Record, and Why Should We Care?" *American Archivist* 57 (Fall 1994): 592–94; and "Re-Discovering the Archival Mission: The Recordkeeping Functional Requirements Project at the University of Pittsburgh: A Progress Report," *University of Pittsburgh Recordkeeping Functional Requirements Project: Reports and Working Papers* (Pittsburgh: School of Library and Information Science, University of Pittsburgh, 1994), 1–33.

[11] David Bearman, *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations* (Pittsburgh: Archives & Museum Informatics, 1994), 148.

[12] Bearman, *Electronic Evidence*, 148.

cation," and "who said what" but may not necessarily use labels. Generally, physical characteristics help to identify archival record types. In fact, physical characteristics may help to identify the archival record type even if content or context cannot be thoroughly understood (e.g., if the record is in a foreign language, or the archival arrangement scheme is not decipherable.) A final point to be made here is that the structure of a record—be it a memorandum, minutes of a meeting, or an invoice—also suggests the specific function of the transaction.

Context defines a record's place within the recordkeeping system from which it emerged. Context is maintained through adherence to the archival principles of both provenance and original order. The principle of provenance holds that "records/archives of the same provenance must not be intermingled with those of any other provenance."[13] (Provenance is the institution or individual that created, accumulated, and/or maintained and used the records in the conduct of business prior to their transfer to an archives or record center.[14]) The principle of provenance appears to have been first codified as *respect des fonds* in an 1841 circular of the French Ministry of the Interior.[15] The principle of original order holds that the "archives of a single provenance should retain the arrangement . . . established by the creator in order to preserve existing relationship and evidential significance."[16] This principle appears to have first been codified in the 1898 work, *Manual for the Arrangement and Description of Archives*, by Dutch archivists Muller, Feith, and Fruin.[17]

Context, perhaps the most critical property of an archival record, is essential for a record to exist as evidence. Too often the property of context is not fully understood, as different provenances are intermingled and/or records are arranged in artificial schemes (e.g., alphabetical or chronological order) that disrespect original order. Records in these scenarios can only exist as "artifacts" simply because their provenancial identity is incomplete and their organic relationship within their recordkeeping system is severed.[18]

As David Bearman points out in relation to data migration, the value of records as evidence is reduced when they are stripped of contextual informa-

[13] Lewis J. Bellardo and Lynn Lady Bellardo, comps., *A Glossary for Archivists, Manuscript Curators, and Records Managers* (Chicago: Society of American Archivists, 1992), 27.

[14] Bellardo and Bellardo, *A Glossary for Archivists*, 27.

[15] Theodore R. Schellenberg, *European Archival Practices in Arranging Records*, Staff Information Paper 5 (Washington, D.C.: National Archives and Records Service, 1939, rep. 1975), 4–5.

[16] Bellardo and Bellardo, *A Glossary for Archivists*, 30–31.

[17] S. Muller, J. A. Feith, and R. Fruin, *Manual for the Arrangement and Description of Archives*, translated by Arthur H. Leavitt (New York: H. W. Wilson Co., 1968).

[18] Glenda Acland, "Managing the Record Rather Than the Relic," *Archives and Manuscripts* 20, no. 1 (1992): 57–63.

tion.[19] Consider the memorandum example independent of context: Its provenance may not be discernible, and surely clues about additional records relevant to its construction will be absent. An accurate interpretation of this record invites a number of immediate questions. What is the provenance of the memorandum? What records supported the creation of the memorandum? Was the memorandum a response to an earlier memorandum, a report, a press release, or general corporate news? Did other memoranda succeed the content in the memorandum under observation? The answer to these questions and the similar questions presented when other types of records are viewed independent of context will only be provided if the contextual property of records is preserved.

Archival Objectives: Evidence, Accountability, and Memory

Understanding the archival properties of content, structure, and context is central to understanding the archival objectives of evidence, accountability, and memory, which constitute a triptych that has emerged from the past decade of discussions and research on electronic records management. These three archival objectives are not mutually exclusive; rather, they build upon each other in a way similar to that of archival properties.

Evidence, the key archival objective, involves documenting transactions—that is, a change in the relationship(s) between an institution(s) and/or an individual(s). Examples of transactions include an invoice documenting a change in ownership of a product; a contract detailing an agreement between any two parties; and a professional or personal correspondence declaring the change in a relationship (e.g., correspondence informing an individual about the termination of his/her job, or—more upbeat—correspondence informing a friend about the birth of a new family member). Preserving context via the principles of provenance and original order is essential to satisfying the evidential objective.

Accountability, as an archival objective, is dependent on the evidential objective. Accountability deals with efficiently accessing evidence in order to verify that an institution or an individual is responsible for an act.[20] Institutions are usually held accountable for their actions; with the exception of illicit or unpopular activities, institutions should want to be accountable for their activities. Individuals, while not accountable as strictly as institutions, are accountable for financial, legal, educational, and even certain health records. Self-employed individuals such as doctors and lawyers need to be accountable

[19] Bearman, *Electronic Evidence,* 146.

[20] Sue McKemmish, "Recordkeeping, Accountability and Continuity: The Australian Reality," in *Archival Documents: Providing Accountability Through Recordkeeping,* edited by Sue McKemmish and Frank Upward (Melbourne: Ancora Press, 1993): 9–26.

in ways similar to those of institutions and, therefore, keep records of their client-related activities.

Accountability enables the institution or the individual to function positively and productively. Consider a collection of records produced by a hospital's radiology departments (X rays, sonograms, and so forth). If the records cannot be retrieved, the corporation is labeled as unaccountable and irresponsible because it can not service doctors or clients, and can only function at a less-than-optimal level. The same scenario exists for an individual who cannot retrieve records (e.g., personal financial reports) required to carry out a specific task, such as trying to purchase a home. On the other hand, an institution or individual under investigation for an illicit activity may desire to be unaccountable. However, history has demonstrated that, at the base of a profitable yet illicit activity, an excellent recordkeeping system often provides incriminating evidence. A case in point is the recent uncovering of the "incriminating" records documenting Philip Morris's testing of tobacco addiction levels.[21] Records that may have contributed to Philip Morris's financial empire may now cause its demise. In sum, our daily news is full of stories of institutional or individual unaccountability that has led to fiscal, societal, and physical harm.

Memory is defined in *Webster's Third New International Dictionary* as the "power or process of reproducing or recalling what has been learned and retained . . . conscious or unconscious evocation of things past."[22] The growing body of literature on the historical understanding of memory—with works by authors such as Patrick Hutton and David Lowenthal,[23]—along with the literature emerging from the field of cognitive science, both provide rich analyses of how this concept is indispensable to individual and societal existence. The archival understanding of memory is very much tied to these analyses because it tells the who, what, where, when, how, and why of an event, activity, or transaction. Recordkeeping systems that preserve provenance along with original order help to satisfy the requirements for institutional or individual memory.

Again, consider the corporate memorandum example. Data following the memorandum's "to" and "from" headers document who wrote it and to whom it was distributed.[24] These data provide information about who's

[21] "The Tobacco Wars: Smoking Documents," *New York Times*, 14 April 1996.

[22] *Webster's Third New International Dictionary of the English Language, Unabridged* (Springfield, Mass.: G. & C. Merriam Co., 1981), 1409.

[23] Patrick Hutton, *History as an Art of Memory* (Hanover, N.H.: University Press of New England, 1993), David Lowenthal, *The Past is a Foreign Country* (New York: Cambridge University Press, 1985), and Lowenthal, *Possessed By the Past: The Heritage Crusade and the Spoils of History* (New York: Free Press, 1996).

[24] Distribution lists for a memorandum do not always accurately indicate who received it, let alone

who in the institution's hierarchy. Data following the "regarding" header (usually "RE:") and the textual body form the memorandum's content, or "what it is about." Together, the who and what data identify where in the institution certain transactions occur, while the "date" data indicate when a transaction took place. Finally, as long as the memorandum's contextual place is preserved in the recordkeeping system, all of these data combined (the who, what, where, and when) provide clues about the how and why a transaction occurred. In short, records and recordkeeping systems must be maintained if the who, what, where, when, how, and why are to be interpreted correctly. Otherwise, context is lost, and the memory objective, as well as the accountability objective, is deficient.[25]

The next section of this article provides a brief overview of the basic NLP information retrieval (IR) framework in which most NLP operations take place.

The Basic Information Retrieval (IR) Framework

Today, the bulk of NLP operations are executed in a basic IR framework, which is composed of three main activities: indexing, query formulating, and matching. A brief explanation of these activities is presented below to help with the final analysis of NLP's applicability to archival properties and objectives.

Indexing

Indexing involves constructing a document representation. A document vector, which is essentially a bag of terms or phrases that represent a document's content, is a commonly used document representation in NLP operations. Traditional IR techniques such as stemming, term frequencies, and weighting, combined with NLP/rudimentary AI processing levels such as syntax and semantics, provide numerous methods for constructing document vectors.

Query Formulating

Query formulating involves transferring a mental query (searcher's information need) to a query representation. A query can be represented by a

who thought it was important enough to read. A memorandum also is often addressed to a group of people such as the "Public Relations Division," which provides no clue as to how many people received the memorandum. At least with a well-designed electronic recordkeeping system, the life cycle of the record can be tracked to the point of knowing who received a memorandum, who kept it, who deleted it, and how it may have been further used.

[25] See Cox, "The Record in the Information Age," for a more in-depth discussion of memory.

single term, a group of terms, a grammatical sentence, a paragraph, or even a full document ("cut and paste" options supported by many of today's graphical user interfaces simplify paragraph and full document querying). A query vector is similar to a document vector in that term weighting and other IR techniques are used in its formulation. A query vector is among one of the most common query representations used in NLP operations. NLP/IR and NLP/rudimentary AI techniques provide numerous methods for constructing query vectors.

Matching

Matching, the focus of the IR process, seeks to compare and match the query representation to the document representation—or the query vector to the document vector, as the vector model is popular with NLP operations.[26] The object is to find the document's text that is similar to, or matches, the query's text. Matching should have a broad interpretation, as it exists on a continuum that ranges from strict textual matching (symbol, word, sentence, etc.) to content matching (meaning or semantics). Matching may also involve a ranked output or relevance feedback: That is, retrieval results (matches) that involve statistical processes can be ranked according to a relevance score. Through relevance feedback processes, the retrieval results can also be used to further enhance or modify an initial query to obtain a more relevant document set.

NLP's Application to Archives

A complete investigation of NLP's ability to index and provide access to archives requires both an in-depth analysis of the various points along the *NLP Continuum* (see Figure 1) and a review of the larger body of NLP research.[27] While these requirements extend well beyond the limits of a single article, NLP's impact on archival properties and objectives can still be explored on a theoretical and somewhat practical level.

NLP and Archival Properties

The overview of NLP and the discussion on archival properties help to illustrate that although NLP is clearly applicable to the archival property of

[26] This IR model review focuses on Basic NLP and some NLP/IR method, but a number of the points highlighted are applicable to more complex NLP/IR operations. For more information on the vector model see Robert F. Korfhage, *Information Storage and Retrieval* (New York: John Wiley & Sons, Inc., 1997).

[27] For an overview of NLP research, see Alan F. Smeaton, "Prospects for Intelligent, Language-Based Information Retrieval," *Online Review* 15 (Dec. 1991): 373–82. See also the bibliography at the end of this article.

content, it has no direct relation to the archival properties of structure and context. (The archival property of context, which involves the principles of provenance and original, is not to be confused with the NLP meaning of context, which involves the manipulation of symbols, words, and other textual structures within a document.) If a record's content is altered by adding or deleting text, NLP will produce different results because the textual representation of the record's content, which also serves as the substance for the NLP operation, has been changed. If, however, a record's structure or context is altered by migrating records, imposing a new organizational scheme on records, or another similar processing activity, an NLP operation will produce exactly the same results—because in any of these cases, the document's content remains the same.

NLP's Strengths and Archival Content

NLP's exclusivity to the archival property of content should not interfere with the realization that it is an extremely powerful operation—one that takes advantage of text stored in electronic form by utilizing the computer's computational capabilities, which surpass human capabilities in speed and consistency. What is so exciting about NLP in the electronic archival environment is that it can contribute to cost-efficient, timely, and consistent indexing; a reduction in human error; and expedient, user-friendly, and exhaustive access. While the following sections briefly highlight several of NLP's key strengths in more of an archival context, it should be emphasized that NLP discussions in the IR and other related literature are also applicable (see the bibliography at the end of this article). The points reviewed are not mutually exclusive and they are not without tradeoffs or weakness.

• *Cost-efficient and Timely Indexing*

NLP offers a cost-efficient means of indexing text, especially when compared to the cost of maintaining an in-house controlled vocabulary specific to an individual archives. That is, NLP automatically manipulates electronic archival text and creates a document representation (e.g., document vector) in a matter of seconds, a method much more expedient and cheaper than that experienced with a human indexer. The cost-effectiveness, however, applies only to indexing time, not necessarily indexing accuracy. Research has demonstrated that automatic indexing can produce favorable results for some types of documents and unfavorable results for others.

In terms of manual indexing, archives can certainly share the cost of developing a controlled vocabulary. In fact, many archival institutions work with subject heading lists and thesauri developed for the library or other

information retrieval communities. Archival vocabulary needs, however, are arguably unique, and the use of jointly or commercially produced vocabulary tools is frequently supplemented by an in-house controlled vocabulary. In short, although the results may not always be favorable, NLP offers a means of indexing archives with discipline specific terminology in an timely manner—archives that might otherwise remain inaccessible because of the high cost associated with traditional indexing operations.

• *Consistent Indexing*

The ability of the computer to be more consistent than the human reveals another NLP strength. Consider the problem of inter-indexer consistency that arises when the work of a single indexer or a group of indexers is inconsistent, even with respect to a single document indexed at different times. This problem is well known and difficult to control because human indexers continuously bring new knowledge, new experiences, new perceptions, and even a host of different moods to their indexing assignments.

With NLP, the problem of inter-indexer inconsistency is nonexistent. An NLP operation will produce the same results, on the same document, every time—even when the indexing is executed at different times. Clearly, an NLP operation is not always as accurate as a human in terms of "understanding" a document's content, but due to the programmed nature of an NLP operation, the results will be consistent. In fact, the consistent indexing ability is also of value when an NLP processing error is detected because it can be automatically cleaned up by improving the algorithm, usually in a short time frame.

• *A Reduction in Human Error*

Associated with indexing consistency issues are human-generated typographical, spelling, and interpretational errors. Undetected, these errors can contribute to lost or inaccessible records that are costly because they take up storage space in electronic recordkeeping systems (albeit storage space is becoming cheaper).

NLP algorithms are free of typographical and spelling errors, although the documents being searched or the query posed may contain them; and some NLP operations claim to be free of, or limited in their production of, interpretational errors, although more research is needed here. NLP errors may have a better chance of being detected and corrected than human-generated errors. This is because NLP errors usually occur system-wide and are pattern-like, giving them a greater chance of being detected, and their clean-up may involve simple algorithmic adjustments. Human-generated errors oc-

cur case-by-case, making them difficult to detect once in the system, and their clean-up, if detected, requires a human indexer's time.

- ***Expedient Access***

NLP offers expedient access when quick information is desired. Suppose someone wants to know whether his/her oil drilling company was ever involved in natural gas exploration in Venezuela. The individual simply needs to type in the query asking:

> Have we [or company name] been involved in natural gas exploration in Venezuela?

Records retrieved in relation to this query may not present a complete picture, but they can quickly verify that the company has actually been involved in natural gas exploration activities in Venezuela: It may also provide an idea about the extent of the company's involvement in this area. Expedient access, as demonstrated here, is invaluable to an institution or individual, especially when the fast-paced deal-making environment of the business world is considered. Despite the fact that the retrieval results will most likely be incomplete, expedient access to institutional or individual records is preferable to having records remain inaccessible because of the high cost of traditional indexing operations.

- ***User-friendly Access***

NLP can support an entirely user-friendly approach to archival access: It permits searchers to enter queries in their own "natural" languages and frees them of the burden of prelearning how to work with a controlled vocabulary or other search strategies that may seem counterintuitive (e.g., Boolean searching). User-friendly searching is clearly an advantage because on-line search-training opportunities do not appear to be a top priority for archival custodians or for the users of archival materials. Perhaps this is a result of the usual immediacy associated with the needs of archival records and the limited time archivists and archival record users have to allocate towards on-line training—even though an initial search-training time investment by both parties may be worthwhile in the long run.

Searching via natural language also includes the benefit of allowing the use of most current terminology. An individual can search with terms representing new concepts, which are not yet part of a controlled vocabulary because of the human processing time requirements. While NLP searching with the most current terminology will not always produce the best results, it still offers a means of access where one might otherwise not exist, again because of the high cost associated with traditional indexing methods.

• *Exhaustive Access*

Exhaustive access involves the assignment of index terms to represent the entire contents of a document. NLP manipulation of a document's full text, or even the text of a substantial abstract, can provide exhaustive access to a document's content. While indexing at this level is clearly impractical if not nearly impossible for the human indexer, this type of access is invaluable with archives, which contain unique informational and evidential content.

Supporting this view is the attention given to "on-the-fly" indexing by David Bearman in the summer 1989 issue of the *American Archivist*.[28] Although Bearman's article does not focus on electronic records, his argument for enhancing access to the rich content of archival materials by increasing the number of access points—even at the expense of consistency (authority control)—is applicable to the electronic archival environment. Exhaustive access can also involve the use of a lexical tool to suggest and assign synonymous and other semantically related terms at the time of indexing or during the query process. For example, a thesaurus may suggest the term "Dividends" for "Payouts" (a use/use for relationship), "Economic growth" for "Growth rates," (an associative relationship), and "Income" for "Life annuities" (a hierarchical relationship).[29]

NLP's Weaknesses and Archival Content

While the application of NLP's strengths to the archival property of content ought to be recognized, so too should the weaknesses, or limitations, of this operation. NLP's general limitations with content retrieval have been well documented in IR and other related literature (see bibliography at the end of this article). A look at several NLP limitations helps to demonstrate the impact this process can have on archival records and recordkeeping systems. These limitations, which are not mutually exclusive, roughly fall into the areas of linguistics, relevancy, processing, and output.

• *Linguistics*

Linguistical ambiguities found with NLP affect the electronic archival environment as they do any IR environment. Some of the most obvious linguistical ambiguities are presented by homographs, morphological processing and word variations, syntax, semantics, and anaphora.

1. *Homographs.* Homographs are distinct words having the same spelling and pronunciation. Consider the sentence, "Our bank is not

[28] David Bearman, "Authority Control Issue and Prospects," *American Archivist* 52 (Summer 1989): 286–99.

[29] Thesaurus examples taken from *ProQuest® Controlled Vocabulary* (Ann Arbor, Mich.: UMI, 1997).

very secure." Does *bank* refer to the financial institution or to the stretch of land at the edge of a body of water? The word could also refer to a container-like object, perhaps in the form of a pig, with a slot at the top for monetary change collection. Distinguishing among homographs can be a simple matter for a human when the context (the NLP sense, meaning sentences within a document) is clear. (See the semantics discussion below.) Establishing NLP context during a computer processing activity, however, presents a difficult challenge.

2. *Morphological Processing and Word Variations.* Morphological processing deals with the grammatical study of word structures, typically roots, stems, and affixes. Word variations such as "child" and "children" or "flies" and "fly" present an obstacle during morphological processing. For example, "child" has a different meaning than "children," and if the "es" is removed from "flies" during the stemming process, "fli" has no meaning at all.

Stemming during morphological processing can lead to other surprises too. Consider the word "representing." Eliminating the prefix "re" and suffix "ing" during morphological processing results in the word "present," which has a number of different meanings, one being a gift, and all of which are very different from the original term "representing." Overall, few NLP operations can account for the numerous word variations that may be misunderstood during morphological processing.

3. *Syntax.* Syntax deals with grammatical rules for structuring language (e.g., placement of verb and noun in a sentence). You may easily understand that the sentences "Jack jumped over the candlestick" and "The candlestick was jumped over by Jack" have the same meaning; however, it is quite difficult to program a computer to understand these compositions.

Another problem with syntax is seen when considering the sentence "The meeting was very time consuming." The computer may have a difficult time here because the activity "meeting," a word also commonly used as a verb, is actually the subject of the sentence. Programming an NLP engine to identify the subject in the sentence "The movie was very time consuming" is much easier because the concept of "movie" as a subject is more concrete than the action word of "meeting." (This example also relates to semantics.)

4. *Semantics.* Semantics deal with contextual meaning. Consider the sentence "Bob saw the client's oil refinery flying over the valley." Was Bob in an airplane flying over the valley and looking down at the oil refinery? Or was he on the ground looking up at the oil refinery flying over the valley?

Words are not always good content indicators; they are dependent on the context in which they are initiated. Despite this fact, the bulk of NLP processing activities operate in what is called a context-free environment. That is, work at the semantic and pragmatic processing levels is an exception because building a semantic knowledge base is a labor-intensive and very costly undertaking for any institution, which of course includes an archives.

5. *Anaphora.* Anaphora is the process of referring back to something that was mentioned earlier in a text. The most common anaphoric expressions include pronouns, but other words and phrases can also be used.[30] Consider the sentence, "The lack of cash flow forced it into bankruptcy." You might immediately understand that the "it" refers to a specific "company" (or other institution), but the computer has a difficult time making this distinction. Language is full of anaphora, making the construction of a knowledge base that can handle this linguistical ambiguity a daunting challenge.

• *Relevancy*

Relevancy, another NLP weakness, is related to the linguistical ambiguities just reviewed. Relevancy, for the purpose of this discussion, involves the mathematical measures of recall and precision, which appear to be the most widely used measures in IR research. These measures are restricted to the topic of a document. There is a growing body of literature that investigates relevancy beyond these traditional measures by looking at utility, user characteristics, record quality, and a number of other factors, which must also be considered.[31]

With respect to the stated interpretation of relevancy, recall is the ratio of relevant documents retrieved compared to the total number of relevant documents in the entire database, and precision is the ratio of relevant documents retrieved compared to the total number of documents retrieved (see Figure 3 for recall and precision formulas). Ideal retrieval results will yield a high recall score together with a high precision score; however, research has

[30] John F. Sowa, ed., *Principles of Semantic Networks* (San Mateo, Calif.: Morgan Kaufmann Publishers, Inc., 1991), 10.

[31] For example, Clare Beghtol, "Retrieval Effectiveness: Theory for an Experimental Methodology Measuring User-Perceived Value of Search Outcome," *Libri* 39 (March 1989): 18–35; Michael B. Eisenberg, "Measuring Relevance Judgments," *Information Processing and Management* 24, no. 4 (1988): 373–89; Stefano Mizzaro, "Relevance: The Whole History," *Journal of the American Society for Information Science* 48 (Sept. 1997): 810–32; *Journal of the American Society for Information Science* 45 (April 1994) (whole issue); Tefko Saracevic, "RELEVANCE: A Review of and a Framework for the Thinking on the Notion in Information Science," *Journal of the American Society for Information Science* 26 (November/December 1975): 321–43; and Linda Schamber, "Relevance and Information Behavior," in *Annual Review of Information Science and Technology (ARIST)* 29, edited by Martha E. Williams (Medford, N.J.: Learned Information, Inc., 1994), 3–48.

$$\textit{recall} = \frac{\text{relevant documents retrieved}}{\text{total number of relevant documents in the database}}$$

$$\textit{precision} = \frac{\text{relevant documents retrieved}}{\text{total number of documents retrieved}}$$[32]

FIGURE 3. Recall and Precision

demonstrated that these two measures generally have an inverse relationship. That is, a high recall score will usually yield a low precision score and vice-versa. Depending on the searching circumstances, a high recall or a high precision score may satisfy the searcher. Even so, systems strive to achieve a balance between these two measures, and a result with an extremely poor recall score or a severely deficient precision score can interfere with satisfying a searcher's information need.

1. *Poor Recall.* Consider the previous query which used the terms "Venezuela" and "natural gas." This query will not retrieve relevant records containing the terms "Caracas" (a city in Venezuela) or other Venezuelan geographic terms, "oil," "petroleum," or other natural gas related terms unless these records also contain the initial search terms "Venezuela" and "natural gas." A lexical tool may help reduce poor recall resulting from this initial search by suggesting terms related to the original query terms; however, an NLP search cannot account for cases in which code words have been used to denote the description of a natural gas exploration project in Venezuela. (The use of code words is not that uncommon for projects that involve outstanding financial or legal risks.) Clearly, relevant records may not be retrieved during an NLP operation simply because their texts do not match the initial user input. The immediate result can yield a poor recall score.

2. *Low Precision.* The flip side of the poor recall problem is lack of precision, which is explained as the retrieval of large amounts of irrelevant records. Again, consider the natural gas exploration in Venezuela query. A basic NLP search with the terms (A)"Venezuela" and (B)"natural gas" can retrieve records that deal with A but not B or records that deal with B but not A. The result is similar to what one experiences with many of today's World Wide Web search engines, which operate with an implicit Boolean "or" operator and present users with many irrelevant records. The use of a Boolean "and" operator can combat this problem, but most untrained searchers are not aware of this option, and their results often yield a low precision score.

[32] Most basic IR and subject analysis text books describe this relevancy equation. For example, see A. C. Foskett, *The Subject Approach to Information*, 5th ed. (London: Library Association Publishing, 1996), 16.

• *Processing*

A third NLP weakness is revealed by the fact that the bulk of NLP/IR operations take place in a preprocessed environment. Most NLP/IR operations involve matching document representations with query representations or document vectors with query vectors. Using query and document representations, as opposed to actual queries and actual documents, makes it impossible for NLP to ever reach the true level of discourse. That is, NLP needs to support a truly dynamic exchange between the actual query and the actual document if discourse is ever to be achieved. The query should react directly with the document and vice-versa. As computer capabilities increase and costs decrease, it seems likely that more NLP experimentation will be seen in dynamic environments.

• *Output*

A final NLP limitation is revealed when the output (retrieval results) are ranked by frequency counts and/or other IR techniques. Ranking methods will often place smaller documents at the end of a ranked retrieval, even though they may be more relevant to the query. To balance the NLP operation, some systems divide larger documents into smaller subdocuments during the NLP processing activity, making for a more fair manipulation of text. This situation, however, does not change the fact that ranked output activities depend on document size, word count, or other related IR algorithms, and are therefore at odds with preserving the context of records in a recordkeeping system.

NLP and Archival Objectives

Given NLP's exclusivity toward the archival property of content, one may wonder how NLP affects the archival objectives of evidence, accountability, and memory.

Evidence

The fact that NLP has nothing to do with the archival property of context results in its incapacity to deal with the archival objective of evidence. Simply put, NLP has virtually nothing to do with ensuring the existence of archival records and recordkeeping systems.

NLP supports a bottom-up approach to indexing and accessing archives.[33] NLP's matching activity involves pulling content information from

[33] Although the spectrums are different, one may draw a parallel between the bottom-up approach

records; via a linking process, records can then be pulled from the context of their recordkeeping systems. This linking activity is really more of a system design issue, not an NLP function, but it is initiated by an NLP search and it illustrates the difficulties that arise when distinguishing between system architecture and system functionalities. Overall, this pulling activity (content from records and records from the context of the recordkeeping system) is detrimental to the integrity of the archival recordkeeping system because it ignores the contextual preservation required for archival records to have value as evidence.

Accountability and Memory

Despite the fact that NLP offers little promise for the evidential objective, it can support the accountability and memory objectives of archives. NLP's ability to aid these two objectives is based on the indexing and accessing strengths found in the archival property of content discussed above.

Consider the earlier query about natural gas exploration in Venezuela. A user with this query can benefit from NLP's strengths such as expedient, user-friendly, and exhaustive access when searching a collection of records. The NLP search probably will not retrieve a complete picture (all the relevant documents); however, any records retrieved can help with the accountability and memory objectives by confirming that the corporation has been involved in natural gas exploration in Venezuela and by providing some information about the nature of its activities in this area. That is, the retrieval results may be able to verify that the oil drilling company is accountable for a number of transactions involving natural gas exploration in Venezuela. The records retrieved can also contribute to the company's memory of its activities in this area.

From this example, it follows that NLP's other strengths, such as cost-efficient and timely indexing, exhaustive access, consistent indexing, and a reduction in human error will also affect the accountability and memory objectives positively. It also follows, however, that the weaknesses associated with NLP's exclusivity towards the archival property of content will affect the accountability and memory objectives negatively.

Again consider the issues associated with the relevant terms "oil," "petroleum," and "Caracas," which did not appear in the original query on

viewed with archival information system access and the bottom-up approach viewed with NLP syntactical parsing, in that both processes begin at the lowest level of the entity being accessed (e.g., an archival collection or a document's text). Conversely, the top-down approach viewed with an archival information system and the top-down approach viewed with NLP syntactical processing begin at the highest level of the collection or the document. See David Bearman, *Archival Methods* (Pittsburgh: Archives & Museum Informatics, 1989), 31–37, for a further explanation of archival information system structure and the bottom-up/top-down approaches to archival access, and Dean et al., *Artificial Intelligence Theory and Practice*, for a discussion of bottom-up/top-down NLP syntactical processing.

natural gas exploration in Venezuela. As previously pointed out, an NLP operation will not retrieve records with relevant terms instead of the initial search terms, unless a lexical tool is used, or unless the searcher conducts multiple queries using the universe of relevant terminology. Clearly, serious consequences may arise if future transactions, either corporate or individual, are based on an incomplete set of records retrieved via NLP or any other process. In a simple sense, it is like making a dish without having added all of the essential ingredients.

Linguistical ambiguities, relevancy problems, absence of a dynamic processing environment, and ranked output problems—all of which are NLP weaknesses associated with the archival property of content—interfere with the accountability and memory objectives. That is, these weaknesses contribute to the construction of incomplete and irrelevant retrieval results, making the complete fulfillment of the accountability and memory objectives impossible. These weaknesses help to confirm the fact that NLP is a content or text manipulator, not a process for preserving records and recordkeeping systems. Moreover, these weakness confirm that NLP cannot exist as the be-all and end-all for indexing and accessing archives. NLP not only has nothing to do with the archival properties of structure and context, but also provides limited support of the archival accountability and memory objectives; NLP also completely fails to support the evidential objective—the true nature of archival records.

Conclusion

It would be impractical to draw an absolute conclusion regarding NLP's potential for archival control because there has been virtually no empirical testing of its capabilities in relation to archival properties and objectives. On the other hand, it would be pointless to invest in an NLP operation thinking it will work equally well for each archival property and for each archival objective. The most sensible approach in adopting NLP for an archival operation is to recognize its exclusivity towards the archival property of content; to take full advantage of its indexing and accessing strengths; and to reduce, as much as possible, the impact of its weaknesses.

In an effort to take full advantage of NLP, archivists need to support systems with a sophisticated linking feature and a mechanism for both bottom-up and top-down indexing and accessing options. This sophisticated linking feature must permit any retrieved record to be viewed within the context of the recordkeeping system from which it emerged. That is, rather than pulling a record from the context of its recordkeeping system, a retrieved record should serve as a means of entry (a link) into its recordkeeping system. Without this improved linking feature, a record retrieved via NLP primarily

acts as a document or artifact, not as a true archival record because it is disengaged from its contextual existence.

With the sophisticated linking feature in place, both the bottom-up and top-down approaches to indexing and accessing archives can be exploited. The bottom-up approach permits NLP's strengths (expedient access, user-friendly access, and so forth) to apply to the entire body of the recordkeeping system. This approach is extremely powerful as it can automatically provide an archival access starting point that may take a tremendous manual effort to find through human processes alone. With the bottom-up approach, however, NLP's weaknesses also operate on a grand scale; tracking the life cycle of the retrieved records via the linking process may present a very timely and complicated process.

The top-down indexing and accessing approach may serve as a vehicle to counter the weaknesses found in the bottom-up approach. With this method, searchers start by selecting an institutional or individual activity, which is represented by a group of records. This approach will not capitalize on NLP's strengths and weakness for the entire body of a recordkeeping system, as is evident with the bottom-up approach. However, it can facilitate a more powerful use of NLP within a subset or a smaller group of archival records because NLP's strengths are easier to exploit and weaknesses easier to control within a smaller group of documents (or, in this case, archival records). Tracking the life cycle of a record is less complicated with the top-down approach than with the bottom-up approach because there is less retrieval with which to work. The drawback of the top-down approach, however, is that the searcher must first select an activity (subset of records) with which to begin the query, limiting the total access options from the onset.

Although more testing is needed to identify the potential for combining the bottom-up and top-down approaches, it seems that allowing a searcher to enter a system either way (bottom-up or top-down) and allowing these two approaches to feed off of each other is the best method for taking full advantage of NLP. Depending on the size and nature of the recordkeeping system, the initial query, and the initial retrieval results, a user could refine a query by switching back and forth between the bottom-up and top-down approaches.

In closing, questions concerning today's archival priorities need to be put forth. Should archivists be spending their limited resources and energy investing in NLP and digitizing archival collections? Does it make more sense for archivists to be testing functional requirements and metadata standards for the preservation and use of records and recordkeeping systems?[34] What

[34] The Recordkeeping Functional Requirements project at the University of Pittsburgh, School of Information Sciences is one project that has taken the lead in identifying functional requirements and metadata for electronic recordkeeping systems. See <http://www.lis.pitt.edu/~nhprc>, which also has links to other sites. See also Wendy Duff, "The Influence of Warrant on the Acceptance

other steps should archivists take to develop adequate electronic recordkeeping systems and intelligently employ NLP?

Archivists can be sure that research will continue at all levels of the *NLP Continuum* presented at the beginning of this article. It therefore makes sense for archivists to monitor and test, when feasible, NLP developments, especially when such developments can enhance access to records in current and developing recordkeeping systems. Perhaps the most critical point to be made, however, is that archivists need to understand the archival record, the fundamentals of archival properties and objectives, and the electronic archival environment so that NLP is not employed with unrealistic expectations.

Appendix—NLP Bibliography

Selected Background Literature on NLP (includes Basic NLP to Full NLP/AI)

Allen, J. *Natural Language Understanding,* 2d ed. Redwood City, Calif.: Benjamin/Cummings Pub. Co., 1995.

Communications of the ACM [Association for Computing Machinery] 39, no. 1 (1996) (whole issue).

Doszkocs, T. E. "Natural Language Processing in Information Retrieval." *Journal of the American Society for Information Science* 37 (July 1986): 191–96.

Dym, E. D., ed. *Subject and Information Analysis.* New York: Marcel Dekker, Inc., 1995 (section 4 includes five chapters on NLP).

Joshi, A. K. "Natural Language Processing." *Science* 253 (1991): 1242–49.

Lancaster, F. W. *Vocabulary Control for Information Retrieval,* 2d ed. Arlington, Va.: Information Resources Press, 1986.

Rowley, J. "The Controlled Versus the Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research." *Journal of Information Science* 20, no. 2 (1994): 108–19.

Salton, G. "Language Analysis and Understanding," Chapter 11, *Automatic Text Processing.* Reading, Mass.: Addison-Wesley, 1989, 377–424.

Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill Publishing, 1983, 257–302.

Smeaton, Alan F. "Prospects for Intelligent, Language-Based Information Retrieval," *Online-Review* 15 (Dec. 1991): 378–82.

and Credibility of the Functional Requirements for Recordkeeping," Ph.D. diss., School of Information Sciences, University of Pittsburgh (1996), and the Metadata Project for the City of Philadelphia: <http://www.phila.gov/departments/records/perp.htm>.

Svenonius, E. "Unanswered Questions in the Design of Controlled Vocabularies." *Journal of the American Society for Information Science* 37 (Sept. 1986): 331–40.

Warner, A. J. "Natural Language Processing," in *Annual Review of Information Science and Technology (ARIST) 22*, edited by Martha E. Williams. New York: Elsevier Science Publisher, 1987, 79–108.

——— "Natural Language Processing: Current Status for Libraries," in *Artificial Intelligence and Expert Systems: Will They Change the Library?* Urbana, Ill.: University of Illinois Graduate School of Library and Information Science, 1992, 194–214.

——— "The Role of Linguistic Analysis in Full-text Retrieval," in *Challenges in Indexing Electronic Text and Images*, edited by R. Fidel, et al. Medford, N.J.: American Society for Information Science by Learned Information, Inc., 1994, 265–75.

Selected Research Articles

Addison, E. R. "Large Scale Full Text Retrieval by Concept Indexing." *Proceedings, 12th National Online Meeting.* Medford, N.J.: Learned Information Inc., 1991, 5–15.

Biswas, G., et al., "Knowledge-assisted Document Retrieval: I. The Natural-Language Interface (and) II. The Retrieval Process." *Journal of the American Society for Information Science* 38 (March 1987): 83–110.

Bonzi, S. and Liddy, E. D. "Testing the Assumption Underlying Use in Natural Language Tests." *ASIS '88: Proceedings of the 51st Annual Meeting.* Medford, N.J.: Learned Information, Inc., 1988, 23–30.

Doszkocs, T. E. "Natural Language Interfaces in Information Retrieval," in *What is User Friendly? Clinic on Library Applications of Data Processing.* Chicago: University of Illinois, Urbana-Champaign, Ill., 1987, 89–95.

Evans, D. A. and Lefferts, R. G. "CLARIT-TREC Experiments." *Information Processing & Management* 31 (May/June 1995): 385–95.

Information Processing & Management 26, no. 1 (1990) (whole issue on "NLP and Information Retrieval").

Metzler, D. and Haas, S. "A Syntactic Filter for Improving Information Retrieval Precision," *Proceedings of the 52nd Annual Meeting of the American Society for Information Science.* Medford, N.J.: Learned Information Inc., 1989, 24–27.

Riloff, E., "Little Words Can Make a Big Difference for Text Classification." *SIGIR'95.* New York, N.Y.: ACM Press, 1985.

Sheridan, P. and Smeaton, A. F. "Application of Morpho-Syntactic Language Processing to Effective Phrase Matching." *Information Processing & Management* 28, no. 3 (1992): 349–69.

Selected Web sites for NLP Research Projects/Products

Natural Language Processing (General NLP site with links to many NLP resources.)

<http://www.cacs.usl.edu/~manaris/ai-education-repository/nlp-resources.html>

CLARIT (NLP software produced by the CLARITECH Corporation. This software is being used for the HELIOS project, which involves the archives of Senator Henry John Heinz III., R-PA.)

<http://www.clarit.com/>

CYC Project (A knowledge base of rules, facts, etc. that supports NLP disambiguation.)

CYC homepage: <http://www.cyc.com/>

For link to NLP see: <http://www.cyc.com/applications.html#www>

SPECIALIST (NLP project based at the National Library of Medicine.)

<http://wwwcgsb.nlm.nih.gov/infotech/nls/>

University of Cambridge NL Group

<http://www.cl.cam.ac.uk/Research/NL/>