DOCUMENT RESUME

ED 406 996 IR 056 374

AUTHOR Greenfield, Rich
TITLE Do We Still Need Controlled Vocabulary? Of Course, We Do! But How Do We Get It: The Roles for Text Analysis Softwares.
INSTITUTION Library of Congress, Washington, D.C. Congressional Research Service.
PUB DATE 16 Apr 97
NOTE 35p.; Paper presented at the CENDI Cataloging Working Group Conference on "The Future of Bibliographic Standards in a Networked Information Environment" (Bethesda, MD, April 16, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Access to Information; *Authority Control (Information); *Automatic Indexing; *Bibliographic Records; *Cataloging; Comparative Analysis; Computer Software Evaluation; *Electronic Libraries; Information Networks; Information Technology; Internet; Library Catalogs; Library Development; *Online Catalogs; Standards; Technological Advancement; World Wide Web
IDENTIFIERS Congressional Research Service; *Document Analysis; Search Engines

ABSTRACT

The author argues that traditional library cataloging (MARC) and the online public access catalog (OPAC) are in collision with the world of the Internet because items in electronic formats undergo MARC cataloging only on a very selective basis. Also the library profession initially isolated itself from World Wide Web development by predicting no real need for universal access, by ignoring large areas of human creativity, and by de-emphasizing "ephemeral" resources. This paper recommends a constructive merger of the best of both worlds--the full text analysis provided by web search engines and the controlled vocabularies found in library OPACs. The Congressional Research Service (CRS) is being used as a testbed to examine relevant techniques. Three of the major text analysis technologies are natural language processing, case-based reasoning, and adaptive learning. As part of "the new OPAC," the Experimental Search System (ESS) is one of the Library of Congress' first efforts to make selected cataloging and digital library resources available over the World Wide Web by means of a single, point-and-click interface. Perhaps even more promising is the idea of using large MARC databases to generate word clusters associated with controlled vocabulary terms and classifications. Six commercial text analysis software products are reviewed in the Appendix, using a comparative table. These tools, many of them associated with major search engine vendors, may support automatic classification and document analysis, thereby increasing cataloger productivity. (AEF)

CENDI Cataloging Working Group

The Future of Bibliographic Standards in a Networked Information Environment

April 16, 1997
Natcher Building, National Institutes of Health, Bethesda MD

ABSTRACT

DO WE STILL NEED CONTROLLED VOCABULARY? OF COURSE WE DO!
But How Do We Get It: The Roles for Text Analysis Softwares

Traditional library cataloging (MARC) does not scale well: more and more items, even those in traditional formats, receive minimal or collection level cataloging, while items in electronic formats are only cataloged on a highly selected basis.

Given that manually produced MARC has largely priced itself out of the Internet marketplace, how can libraries best contribute to organizing the Internet? Are there ways to map the intellectual corpus of MARC - for example, the relationships already established between controlled LCSH vocabulary terms and associated title keywords - to the freetext of the Internet? Several commercial software products, many affiliated with major search engine vendors, claim to have moved beyond fulltext retrieval based on simple word-matching to more sophisticated techniques capable of supporting automatic classification and analysis of fulltext documents equal or superior to that provided by human indexers and catalogers.

Even if these claims are found to be somewhat exaggerated, is there a place for such technologies in the construction of digital libraries? With respect to fulltext documents, can these tools increase cataloger productivity by presenting controlled vocabulary terms for de-selection and by refocusing the cataloger's energies on the editing of machine-generated records and the maintenance of software programs which generate such records?

Several commercial products having potential for improved subject access to fulltext and for automatically or semi-automatically "cataloging" fulltext are reviewed within the context of other existing strategies for indexing and organizing materials on the Internet.

Rich Greenfield, Consultant
Congressional Research Service
Library of Congress
Washington, DC 20540-7241
Tel. 202-707-9104; Fax 202-707-3670

DO WE STILL NEED CONTROLLED VOCABULARY?
The Roles for Text Analysis Softwares

I. WORLDS IN COLLISION

With respect to the library community, we live in a time of worlds in collision: the traditional library world and its highpoint of technology - the OPAC - is currently in collision with (or should I say being eclipsed by?) - the world of the Internet, most particularly in its manifestation as the World Wide Web. Traditionally, libraries have provided access to bibliographic data in electronic form to enable the retrieval of physically formatted objects; the Web, on the other hand, gives direct access to digital objects, if they can be located.

Which of these worlds will survive? Do users even care as long as they get the information they need? I don't think so. But we librarians do care. We *should* care because its our profession and our livelihood, but more importantly, because we have an obligation to our users to represent their interests by building a system which can best serve them.

A. The Triumph of the Containers

Decades ago, the library and cataloging professions seem to have lost sight of the higher purpose of cataloging: *to give access to the full intellectual content and diversity of human creativity.* In addition to describing content, cataloging - before the advent of viewable, computerized formats - necessarily required the description of the physical containers (traditional formats) in which that creativity was packaged - be it paintings, photos, manuscripts, movies, computer software, or whatever - so that these items might be identified, chosen and located. However, in its attempts to be "scientific," i.e., objective, the library profession became a captive of its own rules, mistaking the container for the content and describing the former to the virtual exclusion of the latter (which, of course, is a much more difficult proposition since describing content, as the current "filtering" debate on the Web reminds us, is so subjective).

B. MARC: One Record Fits All

Uniformity in cataloging arose from the historical anachronism of MARC clinging to the format and structure of the printed card where uniformity in treatment made more sense (after all, the cards had to be filed together), and where the emphasis was upon retrieving relevant objects, not upon retrieving relevant content. Why is it, however, that even today practically identical levels of cataloging are given to a small pamphlet and to a CD-ROM containing 500 or 1,000 such pamphlets? *Because librarians still see themselves as cataloging containers, not content, and because in the world of MARC all containers are more or less equal!*

This historical lack of *proportionality* in cataloging has continued despite the computer's ability to dramatically expand and vary metadata according to the type of item being described: things

that are "bigger" (that hold more information, e.g., databases) could have been getting bigger (i.e., more detailed) cataloging, including more controlled vocabulary terms. True, the size of the container usually doesn't matter to the user, but should encyclopedias be cataloged in the same way as books? The overemphasis on describing packaging (object retrieval) has had a direct effect on reference service with many reference librarians asking patrons, "Are you looking for articles or books on your subject?" when, more often than not, patrons are looking for neither. Nor are they looking for "subjects" per se. What they are really looking for are solutions to problems.

Ironically, it is only because of the often despised (by librarians) commercialization of the A&I services - which moved into an indexing vacuum left by libraries - that users consulting bibliographic metadata in their decision-making process now have a better sense of what an article is about than what a book is about - even though in terms of overall content, the article is much less substantial (one hesitates to say "significant") than the average book! Today, faced with a world in which the containers - and even the content - are fluid, librarians can no longer cope - their rules no longer apply. The profession is in crisis.

II. A PROFESSION IN CRISIS

Why was Web development so independent of the OPAC? In my opinion, the profession made several mistakes: first, because it based its work on the holdings of individual libraries, not the universe of human creativity, it predicted no real need for and possibility of universal access; secondly, it chose to ignore - largely because of its inability to scale its overly rigid, rule-based cataloging - huge areas of human creativity (e.g., photography, fiction, poetry, songs, etc.); and, thirdly, it belittled as ephemeral resources that its workflow was too slow to handle (e.g., case law, electronic pre-prints, email ,etc.) As a consequence, entire professions - namely, the sciences, law and business - were forced to develop their own mechanisms for communicating and storing information. In short, librarians as a profession played it safe by sticking to the most stable of information delivery formats, printed books and serials.

A. Islands of Information

Besides the gradual and largely independent development of TCP/IP networking, there were probably many other good reasons for the early isolation of library communities from the web; for example, it was both an ethical standard and a practical mechanism of quality control that, until recently, the professional bibliographer would never include a work in a published bibliography that had not been *physically* examined. Hence, the Internet was not perceived as a tool of value in the construction of bibliographies, nor foreseen as the source for the automatic and semi-automatic construction of bibliographies that it is has become today. Unfortunately, such narrowness of vision left libraries for too long as islands of information and this legacy leaves the profession vulnerable to the broader and potentially more in-depth indexing and

efficient classification which web-based search engines are just beginning to provide.

B. Universal Awareness (and Access) Denied

It has always been the case that serious researchers usually want to know what exists, even though they may consult what is available locally, before deciding how much they are willing to pay, how much time they are willing to spend, and how far they are willing to travel to access materials at a distance. In basing cataloging on local holdings, librarians left, for the most part, the creation of reference tools - *which do attempt to encompass the universe of human creativity* - to the private sector, even though computers could have allowed for the automatic generation of printed reference tools from underlying OPAC databases.

As any reference librarian knows, even with the advent of online union catalogs, it is the commercial directories and databases of serial articles, CD-ROMs, software, videos, organizations, even books, etc. that now define the universe of knowledge, not the OPAC. With the exception of major research libraries and some state systems which are site licensing commercial resources, reference products are now going online in consumer online and web-based services, but not in most OPACs. Even when mounted in a Campus Wide Information System (CWIS), commercial databases and online reference tools have yet to be tightly integrated with the OPAC, though this may change (e.g., the CIC Reference Shelf project).

III. AVOIDING COLLISION

What is suggested in this paper are some relatively simple mechanisms to prevent these two worlds from violently colliding or worse, simply bouncing off one another and proceeding it different directions. Instead, a constructive merger of the best of both worlds - the fulltext analysis provided by web search engines and the controlled vocabularies found in library OPACs is recommended in this paper.

The work of the Congressional Research Service is being used as a testbed to examine these techniques. The Congressional Research Service (CRS) consists of some 800 federal employees, about a fifth of the Library of Congress work force, who are solely dedicated to meeting the information needs of the Congress. Among other responsibilities, CRS builds and maintains - with the assistance of Information Technology Services - legislative bill tracking systems, most recently LIS (the Legislative Information System), which is derived from its public predecessor, THOMAS at http://thomas.loc.gov.

A. The Problems Faced by CRS

Some of the major files in these two systems (LIS and THOMAS) include databases of Bill Summary and Status information prepared and updated by the Bill Digest Section of CRS. Bill

Digesters create written digests for the 5-6 thousand bills introduced in Congress each year and also assign controlled vocabulary terms to these digests from the Legislative Indexing Vocabulary Terms (LIVT) thesaurus, a collection of over 10,000 descriptors for the field of public policy literature.

Like many other organizations, the problems CRS faces stem from the advent of "webtime," wherein traditional workflow speeds are perceived by users as agonizingly slow. As you know, in webtime, a year equals a month, a month equals a week, a week a day; a day an hour; and an hour becomes a minute. User expectations have been raised by web-based push technologies so that the preparation of bill digests and assignment of indexing terms, which now takes as little as a day or two or as much as a month or two, depending on the legislative agenda, is no longer acceptable.

In search of solutions, the Congress and CRS are looking into the SGML encoding of data at its point of origin; the installation of new, SGML-aware document management and workflow systems; and the potential applications for text analysis software products. It is this last category that will be reviewed here.

B. The Text Analysis Solutions Being Investigated by CRS

The term "text analysis" actually suffers from the disease for which it purports to be the cure: the technology is poorly and loosely defined and it is hard to imagine that computer analysis would give it any greater degree of clarity. Text analysis software products claim to represent a natural evolution from Boolean and relevancy-ranked search engines to greater degrees of content analysis. [see DR-Link chart on p. 22] They rest their claim upon several overlapping technologies which include word frequency lists at one end of the spectrum, adaptive learning techniques and case-based reasoning somewhere in the middle, and true natural language processing at multiple levels of linguistic analysis at the other extreme.

Three of the major technologies are: **Natural Language Processing** (NLP), defined as "a full range of computational techniques for analyzing and representing naturally-occurring text;"**Case-Based Reasoning** (CBR), which has been defined as a techniques or "adapting old solutions to new demands" by comparing a current question or problem to a library of past answers or solutions and interactively guiding the user through an iterative search process; and **Adaptive Learning**, which has been defined as an iterative process with automatic feedback loops, often built upon a "query-by-example" model. The most successful text analysis products seem to have chosen a strategy of layering these technologies one upon the other. [see DR-LINK "Synchronic Model of Language" diagram on p. 22 and the DR-LINK description p. 17]

Interestingly, the same technologies for natural language processing and querying of fulltext can be used (and are being used) for the filtering and routing of documents - now called "push/pull"

services, but traditionally known to librarians as "saved searches" and "SDI services."

The CRS priorities for text analysis software applications are fourfold:

1) machine-assisted summarization (or abstracting) of bills for the creation of preliminary bill digests;

2) machine-assisted assignment of LIVT (Legislative Indexing Vocabulary Terms) to bills and bills digests

3) the provision of computerized, individually customized SDI services which can filter and merge information streams from internal and external databases and deliver metadata linked to digital objects to the analyst's desktop; and,

4) natural language querying techniques for fulltext that go to the level of syntactical and semantic analysis.

IV. LIBRARY APPLICATIONS

Outside of CRS, there are parallel Library Services application which might be of value to libraries generally. These include the automatic or machine-assisted assignment of LCSH (Library of Congress Subject Headings).

A. MARC Cataloging

Machine-assisted assignment of subject headings would be more difficult than bill analysis and assignment of LIVT terms since the majority of LC acquisitions are not machine-readable (i.e., full-text, the major exception being some items in the CIP -Cataloging in Publication-program).Though past attempts have not proved very successful because of the small amount of text available in the average MARC record, there is still some residual interest in the library community in imputing subject headings based on key words in titles and other fields in the MARC record.

For non-digital works of non-fiction, LCSH terms could be automatically assigned using traditional relational database keyword matching and relevancy-ranked search engines with well-defined threshold criteria. Better yet, *adaptive learning techniques* could be applied to a large collection of MARC records (the larger the better) to create a semantic net of the relationships of significant words in specific MARC fields, namely title and notes fields, though perhaps author and publisher fields might also provide relevant information for disambiguation purposes. Suggested subject headings would then be edited - either approved or deselected - by subject catalogers. Similar techniques might be used to automatically generate LC classification (call)

numbers. These might bring about modest gains in cataloging productivity and in reducing somewhat the cost of cataloging.

B. Web Cataloging

The major absurdity of listserv discussions - where librarians beat their chests and swear they can bring order to the Web by cataloging it using traditional means of cataloging - is that these same librarians, most of whom are doing copy cataloging anyway, cannot keep up with their current workloads. They have been unable to get traditional formats under bibliographic control without resorting to shortcuts like minimal level and collection level cataloging. Where will they find the time to catalog 50 million existing Webpages? And at what cost?

Even if all 20,0000 catalogers in the country devoted half a day every day to doing original cataloging for five Websites, it would take a full year for them catalog the first fifty million sites, by which time there would be millions of new sites and millions of old sites which are in need of updating. Because of it incredible rate of growth, any attempt to manually catalog the entire Web is hopeless, even without the problems of broken links and changing site content. It is irrational to believe that the OPAC can subsume the Web, but there are many simple techniques available for merging the traditional OPAC and the Web. The first steps are the obvious ones: putting OPACs on the Web, enriching their internal links, MARC-ifying other databases on the Web, and finally linking outward from individual MARC records to related resources on the Web.

V. THE NEW OPAC

The battle for tomorrow's "next generation" OPAC is being waged on the Web today. Even without positive cash flow, feature-rich web search engine companies have got the development momentum, the venture capital, and the competitive drive to risk abandoning some users in the short run in order to gain more users in the long run as they repeatedly reinvent themselves with the new programming languages like JavaScript, Java, and ActiveX. Most importantly, these search engines exist in "webtime."

The fate of the OPAC, like the fate of the library to which it is tied, depends on how much it can change, how much these two institutions can open themselves to the universe of information beyond the library walls. Users want and need to move effortlessly in their searching from the OPAC to the Web and vice versa. A new generation of OPACs has emerged to meet this challenge and includes Eureka on the Web (RLG) at http://www.rlg.org; Melvyl on the WWW (University of CA) at http://www.melvyl.ucop.edu/, and ESS (LC) at http://lcweb2.loc.gov/resdev/ess/

A. LC's Experimental Search System (ESS)

As explained in its hyperlinked help, the Experimental Search System (ESS) is one of the Library of Congress' first efforts to make selected cataloging and digital library resources available over the World Wide Web by means of a single, point-and-click interface. The interface consists of several search query pages (Basic, Advanced, Number) and a several search results pages (an item list of brief displays and an item full display), together with brief help files which link directly from significant words on those pages. By exploiting the powerful synergies of hyperlinking and a relevancy-ranked search engine (InQuery from Sovereign Hill), ESS developers hope ESS will provide new and more intuitive ways of searching the traditional OPAC (Online Public Access Catalog). [see sample screens p. 24-26]

The functionality of ESS includes the expected *intra-links* between MARC records using the hyperlinked fields that are common in web-based OPACs, i.e., subject heading and author name links. In addition, though, other major non-MARC databases (e.g., American Memory, THOMAS' Legislative Files, etc.) have been MARC-ified to create MARC format records for each item in these databases. The process of MARC-ification of disparate data files and the centralized storage of the resulting records in an OPAC can provide cross-file search functionality which, though far from perfect, has been enthusiastically welcomed by users.

Finally, ESS references web resources that LC may or may not control by creating *inter-links* between MARC records and web resources, wherever they may be located. Concentrating on the most stable of webpages (i.e., not individual pages), these links include thoses from the MARC 856 field to fulltext objects (ASCII, HTML, and PDF books; fulltext legislation; films, etc.); from publisher names to publisher homepages; and, still to come when appropriate, author/subject links to DejaNews and other Usenet archives; to listserv archives; to corporate homepages; to bookstores; and to specialized search engines.

B. Merging the OPAC and the Web

Perhaps even more promising that these linking technologies is the idea of using large MARC databases to generate word clusters associated with controlled vocabulary terms and classifications, i.e., semantic nets. MARC records can be parsed into noun phrases from title and notes fields and associated with specific LCSH or MESH terms and LC or Dewey classification numbers. These clusters of meaning can then be used, through relatively simple matching techniques, to screen the Web, generating and attaching appropriate subject headings and class numbers to fulltext Web documents and document fragments.

Once clustered, the semantic net would allow a simultaneous search of one or more OPACs and the web - of the smaller, bibliographically controlled local universe, and the immense, largely un-controlled Web universe. Being associated with specific controlled vocabulary terms, results

would lend themselves to visualization. It is only a matter of time before major search engine vendors access large MARC database to create semantic nets for improved subject access to unindexed fulltext via their search engines.

The consequences are bi-directional: the OPAC searches are already launchable from within Web search engines; and, Web searches will soon be launchable from within OPACs. Thus, users will launch Web searches to find controlled vocabulary terms in OPACs and then use the controlled vocabulary to launch new Web searches. Every MARC record can have multiple searches (and services) associated with it, e.g., all records with Shakespeare as a subject heading could be linked to Shakespeare listserv archives; MARC records of family genealogies can link to major phone directories or automatically conduct a family name (with its appropriate variations) search, or as an intermediate step, just link to an associated webpage with the links to Web genealogy resources. Particular Web search engines will be associated with specific categories of subject headings, depending on the search engine's domain, e.g., business, law, medicine, etc.

In its related SCORPION project <URL:http://purl.oclc.org/scorpion>, OCLC is already experimenting with some of these techniques through the automatic assignment of metadata based on text analysis of webpages (with decidedly mixed results as detailed in the recent Web4Lib "metadata" thread). Unfortunately, with several notable exceptions like researchers at OCLC, few library and information science researchers or academics are active in the field of text analysis. Only a handful of research libraries do conduct limited research and development projects in this area, but, at the same time, there are already many sophisticated commercial products on the market in use for classifying and summarizing fulltext.

Out of the dozen or more products available (and the dozens more under development),just six have been chosen for review in this document: Convectis, CBR Content Navigator, LinguistX, ConText, InClass, and DR-LINK.

VI. ROLES FOR TEXT ANALYSIS SOFTWARES

CRS research into the usefulness of commercial text analysis software packages has just begun. In the context of the product descriptions and bibliographies provided in the appendices, revisiting initial CRS priorities - machine-assisted summarization of bills for the creation of preliminary bill digests; machine-assisted assignment of LIVT to bills; provision of computerized SDI services to the Congressional staffers desktop; and supporting natural language querying techniques for fulltext retrieval - reveals that this category of product, i.e., text analysis, claims to do all of these things, though not any single product necessarily does them all or does them all necessarily well.

On the basis of a simple literature review, Convectis and ConText would seem to be the most mature products in terms of machine summarization, though DR-LINK may have a superior

technology that in the end will provide better results; for assigning controlled vocabulary, Convectis and ConText again are already in commercial production, while InContext and DR-LINK show a lot of promise; for the routing of filtered information to the desktop, InRoute and Convectis have successful, large scale applications in place; and, finally, for natural langauge querying, DR-LINK and, within certain domains, CBR Navigator, would seem to provide clearly superior technologies. LinguistX is in use with one and perhaps more of these products.

VII. CONCLUSIONS

The next step in the product evaluation process will be to pursue in-depth presentations by these and perhaps other vendors, solicit demonstration versions and/or passwords for web-based access to working prototypes and production systems, and conduct site visits where these products are already installed and in production. Ideally, legislative data from the THOMAS system can be processed and tested with leading products in head-to-head tests, both against each other and existing in-house retrieval systems.

It is clear from preliminary tests that vendor representations of the wonders of their text analysis softwares usually exceed the actual results when used on the client's data *without human intervention.* For example, see the machine-generated abstract of this paper produced by Convectis (p. 27). And yet, these automatic results may prove adequate to meet immediate user and management needs: for abstracts or summaries, until higher quality can be produced by humans; for assignment of controlled vocabulary terms, to expedite overall cataloging workflow; etc. No doubt extensive "tweaking" or "tuning" can often dramatically improve performance

As a final comment, the social aspects of integrating such technologies into existing workforces should not be ignored. Expect to receive some level of resistance from staff - not just those whose daily routine may be modified, but from librarians, especially professional searchers and catalogers. Just the introduction of relevancy-ranked search engines (e.g., THOMAS, LIS and ESS using InQuery), particularly when applied to MARC records, has created considerable controversy.

Two valuable articles with insight into these issues outline the initial discomfort felt by an online searcher using a non-boolean system (a "loss of control," no doubt similar to that felt by many catalogers at the prospect of the automatic assignment of controlled vocabulary terms), both by Susan Feldman, appear in the October 1994 issue of Searcher: The Database Magazine for Professionals and in the November 1996 issue of ONLINE, "Comparing DIALOG, TARGET, and DR-LINK." The latter article can be also be found online at:
http://www.onlineinc.com/articles/onlinemag/feldman9610.html

Aptex's Convectis http://www.aptex.com

Aptex Software, Inc. is a recently created (1996) division of HNC. HNC itself was founded in 1986 by Robert Hecht-Nielson, a leading figure in the commercialization of neural networks, and Todd W. Guschow, who worked together with Hecht-Nielson at TRW's neurocomputing R&D program. A "content mining" technology, Convectis uses neural network and content vector analysis techniques to automate document analysis and categorization. Convectis first analyses word relationships, then assigns categories to documents, and then provides an opportunity to improve categorization though human feedback. [see Aptex diagram page]

Its claimed advantages are that it learns from examples and is therefore easy to get up and running and it has a low cost of ownership because no dictionary or thesaurus maintenance is required (nor is their a rule-base to construct and maintain). It is advertized as ideal for dynamically-changing topics, as scaling well, and as language independent.The underlying insight upon which Aptex software relies is that of vector analysis, based on the work of Gerard Salton, which assumes that words with similar meanings, when analyzed statistically in relation to the words with which they appear, have similar positions in vector space. [see Aptex diagram p. 28]

Convectis takes this a step further by assuming that documents on similar topics also have what they call "Context Vectors" pointing in similar directions. The distance between Context Vectors can be represented mathematically and used to calculate the degree of similarity between documents. The obvious weakness in such a system is that disparate domains can have, historically, entirely separate vocabularies for talking about the same or similar things and Convectis would have difficulty analyzing them as "similar." Conversely, different domains may use the same or similar vocabularies but with distinctly different meanings. This, too, would tend to confuse Convectis. Although human feedback ("tuning") is optional, it can be used to disambiguate vocabularies when problems of overlapping domains arise.

On the input side, Convectis can learn by example from unstructured text. On the output side, it can produce variable length document summaries and keywords which could be linked to a controlled vocabulary. [see p. 27 for the Convectis output of this paper] For purposes of routing and database quality control, Convectis provides duplicate document detection.Convectis has been used by InfoSeek as an "intelligent librarian" to categorize "millions" of WWW pages for its UltraSeek website.

Bibliography:

CONVECTIS: A Context Vector-Based On-Line Indexing System by Robert Sasseen, Joel L. Carleton, and William R. Caid

Inference's CBR Content Navigator http://m5.inference.com

Based in Novato, California, Inference was founded in 1979 and has about 200 employees more or less equally distributed between the US and abroad. Inference's CBR (CBR = Case-Based Reasoning) products provide a "common platform" for the search and retrieval of unstructured information using a patented case-based reasoning system that integrates with a rule-based reasoning system.

Relying upon a technique of comparing a current problem to a library of known problem solutions, natural language must first be input through a simple forms-based interface to build case bases for resolving problems or selecting resources. CBR Express Generator can automate this process by taking any set of documents and immediately creating a case base for access to those documents by statistically analyzing to provide a summary, based on "statistically relevant" phrases in each document, and by generating a set of questions that will help to distinguish between similar documents.

Using Case-Point, an end-user case base search and retrieval application, users can retrieve documents by answering a series of questions rather than conducting Boolean searches. CBR is particularly useful where rules (expert systems) or algorithms (neural nets) are difficult to create and maintain, but where correct solutions are to problems are available, such areas as help desk operations, technical support, and reference interviews, but it would seem to require a highly-defined domain and a history of problems (queries) and documented solutions (answers).

A case consists of a title, a description, an optional multimedia annotation (voice, graphics, etc.), a series of questions and answers, and an action or multiple actions. Questions, which can be assigned weights by their authors, are used to gather information and refine the search. Actions are the pieces of information or the solution to the problem being searched. An intuitive dialogue-like approach guides users by posing a series of question designed to disambiguate and narrow the query.

Inference's CBR products include a document summarization module and have been integrated with Verity's TOPIC search engine. They support Oracle, Sybase, Microsoft SQL Server, Informix, BB2/2 and RAIMA Dta Manager (RDM).

For a sense of how well Inference's search engine categorizes websites, go to:

http://m5.inference.com/ifind/ifind.cgi

and type in a search such as "text analysis" or "natural language processing."

inXight's LinguistX http://www.inxight.com/products/linguistx/overview.shtml

Spun off in 1996, InXight is a Xerox New Enterprise Company whose LinguistX product claims to be "based on 15 years of research in linguistics at Xerox research centers in Palo Alto, California (PARC) and Grenoble, France (RXRC)" and "the fastest, most compact and most flexible collection of linguistic software components available for licensing today." Built upon Xerox Linguistic Technologies (XLT - see bibliography below), LinguistX is primarily licensed to software and search engine developers. It contains advanced natural language processing components that include modules for automatic document summarization, information extraction and morphological analysis. It is currently used by several major search engines, both on the Net and off (e.g., InfoSeek, Verity).

The two basic modules which make up LinguistX are a Document Analyzer and a Query Analyzer. The latter contains a linguistic transducer which does document analysis and word morphology (a more advanced form of stemming). LinguistX indexes the root concepts of words, not the words themselves, and can identify their use as nouns, adjectives or verbs; thus, relevance ranking can be based on more than the matching of identical words.

Features include: *tokenizing* (separation of documents into sentences and individual words); *stemming* (swam/swim/, peut/puisse) - not primitive "tail chopping"; *morphological analysis* (identification of the grammatical features of a word); *tagging* (building on morphological analysis by choosing part-of-speech categories); *morphological inflection and generation* (which can expand a limited query vocabulary; the inverse of *stemming analysis*); *summarization* (into key phrases and extracted sentences); *language identification* (up to eleven languages).

The XLT Summarizer "automatically examines the content of a document in real-time to identify the document's key phrases and extract sentences to form an indicative summary, either by highlighting excerpts within a document or creating a bulleted list of the documents key phrases. LinguistX might be characterized as aseries of NLP modules, rather than a complete system. Most of its clients seem to use it in such a context, i.e., in conjunction with other natural langauge processing

Bibliography:

Xerox PARC's InXight, The BusinessTech tech feature, March 1997
http://businesstech.com/feature/btinxight9703.html

Xerox Linguistic Technology (XLT)
http://www.xsoft.com/XSoft/lexdemo/xlt_welcome.html

Oracle's ConText http://www.oracle.com/products/oracle7/oracle7.3/html/context_opt.html

ConText is an advanced text retrieval technology with language analysis services which "based on natural language processing technology, can automatically summarize and profile text - all via the same integrated SQL interface - for intelligent searches of very large-scale text databases." It is advertized as an option for Oracle7 Release 7.3 Server Among its text retrieval features are: exact word/phrase searching, multilingual stemming, proximity searches, relevance ranking, boolean logic, wild card searching, term weighting, thesaurus support , fuzzy searching and stop lists.

Its two major linguistic features are: a) text extraction and classification - the ability to get the major themes, or ideas, discussed in a document [available for English only]; and, b) text summarization - the ability to get an automatic "gist" or reduction of a document. Theme indexes allow ConText to find documents about themes even when the themes do not appear as actual words in the text. Theme summaries are collections of paragraphs that best represent that particular document theme. Themes are limited to 16 per document, regardless of document size.

ConText provides an API (Application Programmers Interface) that allows developers to add these layers of language-processing software to the tasks performed by their traditional text retrieval and document management products.

Bibliography:

Developing Applications with the Oracle ConText Option - An Oracle White Paper, October, 1996. 16 pages
http://www.oracle.com:81/products/oracle7/oracle7.3/html/devap_w.pdf

Managing Text with ConText Option to Oracle Universal Server - An Oracle White Paper, March 1997. 17 pages

Oracle ConText option 2.0 Data Sheet (PDF)
http://www.oracle.com:81/products/oracle7/oracle7.3/pdf/46871_21372.pdf

Oracle ConText: Text Looms as the Next Frontier in Information Management, by Timpthy O'Brien, April 1996. 9 pages
http://www.oracle.com:81/products/oracle7/oracle7.3/html/context_seybold.html

Text-Enabling Web Applications with Oracle ConText Option
http://textserv.us.oracle.com/oco/webapp.htmlxt

Sovereign Hill's InClass http://www.sovereign-hill.com/

A customizable add-on to the InQuery search engine which has yet to be released as a shrink-wrapped module, InClass uses adaptive learning techniques for processing large sets of data and manually assigned codes in order to learn how to assign codes to subsequent documents. For example, it has been used to assign one of 10,000 primary ICD-9-CM diagnostic codes to patient discharge summaries. Codes have also been automatically assigned to patient/doctor encounter notes.

Another InClass approach assigns to a document as many categories as may be applicable from a choice of categories, while a third approach puts documents in bins or simply assigns a score (1-6 for example) to a document. This approach has been used with essay answers to SAT test questions, successfully (with about the same accuracy as human graders).

Related InQuery modules include.

InRoute: a filtering and routing system that delivers the relevant, real-time information to end user with unique and constantly changing information requirements. Through adaptive learning techniques, InRoute continuously analyzes and adapts its search to pinpoint relevant information. InRoute simultaneously extracts information from multiple heterogeneous and multi-lingual sources for desktop delivery.

InFinder: a contextual thesaurus for users who need to expand the scope of their queries to a concept search level. "InFinder pulls in highly relevant documents that a traditional thesaurus would miss and eliminates documents that a standard thesaurus might erroneously include."

InQuery itself is described by Sovereign Hill as "A distributed information retrieval system that accurately locates and delivers both structured and unstructured information residing on intranets, the Internet and the extended enterprise. Consistently ranked by independent tests to have the highest precision and recall in the industry, the InQuery system was designed to accommodate any data type (text, image, audio and video) in distributed, heterogeneous computing environments. "

Bibliography:

Automated Classification of Radiology Reports by David Aronow and Fangfang Feng. CIIR IAB November 18, 1996.

Oracle's ConText http://www.oracle.com/products/oracle7/oracle7.3/html/context_opt.html

ConText is an advanced text retrieval technology with language analysis services which "based on natural language processing technology, can automatically summarize and profile text - all via the same integrated SQL interface - for intelligent searches of very large-scale text databases." It is advertized as an option for Oracle7 Release 7.3 Server Among its text retrieval features are: exact word/phrase searching, multilingual stemming, proximity searches, relevance ranking, boolean logic, wild card searching, term weighting, thesaurus support , fuzzy searching and stop lists.

Its two major linguistic features are: a) text extraction and classification - the ability to get the major themes, or ideas, discussed in a document [available for English only]; and, b) text summarization - the ability to get an automatic "gist" or reduction of a document. Theme indexes allow ConText to find documents about themes even when the themes do not appear as actual words in the text. Theme summaries are collections of paragraphs that best represent that particular document theme. Themes are limited to 16 per document, regardless of document size.

ConText provides an API (Application Programmers Interface) that allows developers to add these layers of language-processing software to the tasks performed by their traditional text retrieval and document management products.

Bibliography:

Developing Applications with the Oracle ConText Option - An Oracle White Paper, October, 1996. 16 pages
http://www.oracle.com:81/products/oracle7/oracle7.3/html/devap_w.pdf

Managing Text with ConText Option to Oracle Universal Server - An Oracle White Paper, March 1997. 17 pages

Oracle ConText option 2.0 Data Sheet (PDF)
http://www.oracle.com:81/products/oracle7/oracle7.3/pdf/46871_21372.pdf

Oracle ConText: Text Looms as the Next Frontier in Information Management, by Timpthy O'Brien, April 1996. 9 pages
http://www.oracle.com:81/products/oracle7/oracle7.3/html/context_seybold.html

Text-Enabling Web Applications with Oracle ConText Option
http://textserv.us.oracle.com/oco/webapp.htmlxt

Sovereign Hill's InClass http://www.sovereign-hill.com/

A customizable add-on to the InQuery search engine which has yet to be released as a shrink-wrapped module, InClass uses adaptive learning techniques for processing large sets of data and manually assigned codes in order to learn how to assign codes to subsequent documents. For example, it has been used to assign one of 10,000 primary ICD-9-CM diagnostic codes to patient discharge summaries. Codes have also been automatically assigned to patient/doctor encounter notes.

Another InClass approach assigns to a document as many categories as may be applicable from a choice of categories, while a third approach puts documents in bins or simply assigns a score (1-6 for example) to a document. This approach has been used with essay answers to SAT test questions, successfully (with about the same accuracy as human graders).

Related InQuery modules include.

InRoute: a filtering and routing system that delivers the relevant, real-time information to end user with unique and constantly changing information requirements. Through adaptive learning techniques, InRoute continuously analyzes and adapts its search to pinpoint relevant information. InRoute simultaneously extracts information from multiple heterogeneous and multi-lingual sources for desktop delivery.

InFinder: a contextual thesaurus for users who need to expand the scope of their queries to a concept search level. "InFinder pulls in highly relevant documents that a traditional thesaurus would miss and eliminates documents that a standard thesaurus might erroneously include."

InQuery itself is described by Sovereign Hill as "A distributed information retrieval system that accurately locates and delivers both structured and unstructured information residing on intranets, the Internet and the extended enterprise. Consistently ranked by independent tests to have the highest precision and recall in the industry, the InQuery system was designed to accommodate any data type (text, image, audio and video) in distributed, heterogeneous computing environments. "

Bibliography:

Automated Classification of Radiology Reports by David Aronow and Fangfang Feng. CIIR IAB November 18, 1996.

TextWise's DR-LINK (Document Retrieval Through Linguistic Knowledge) at http://www.mnis.net/

Based in large part on work by Elizabeth Liddy and Michael L. Weiner done with Advanced Research Project Agency (ARPA) funding under Tipster, DR-LINK is designed to automate the process research librarians use in transforming the information needs stated by users into effective online queries. The underlying concept is that text retrieval should be at the conceptual level, not the term level. Commercialized through TextWise, Inc. and marketed by Manning & Napier, DR-LINK operates at many levels of linguistic knowledge: *morphological* (smallest meaningful parts of words), *lexical* (the words themselves), *semantic* (meanings), *syntactic* (word order), *discourse* (context) and *pragmatic* (common sense).

DR-Link performs a "staged processing" of documents which reflects the modular structure of its development. These stages include preprocessing where texts are divided into subtexts and parts of speech tagging added (Preprocessor); text structuring in which clauses or sentences are tagged with annotations identifying them in terms of source, time and intentionality as main event, expectation, or consequence (Text Structurer); subject field coding whereby each word in a text is tagged with a disambiguated subject code and each document /unit is represnted as a vector of all the subject field codes of words in that document/unit (Subject Field Coder - SFCoder).; a V-8 SFC matcher that combines the annotations of the Text Structurer and the SFCoder to capture the "discourse meta-components" in a document; proper noun interpretation classifying proper nouns into one of 37 categories (e.g., organization, country, company, etc.) and expansion of nouns into their appropriate hierarchies, e.g., European Community to all member countries (Proper Noun - PN - Interpreter); matching of complex nominal constructs like "debt reduction," "campaign financing," and "electronic theft." (Complex Nominal Phraser); a sublanguage grammar relying on linguistic constructions to recognize and extract the logical combination of relevancy requirements in a users's query (Natural Language Query Constructor); and several other layers of cumulative processing, all of which permit DR-LINK to accept ambiguous, complex natural language queries which it can translate into precise Boolean representation of user relevance requirements. [see page 29 for a diagram]

Bibliography:

"Comparing DIALOG, TARGET, and DR-LINK" by Sue Feldman, ONLINE, November 1996. At http://www.onlineinc.com/articles/onlinemag/feldman9610.html

"Document Retrieval Using Linguistic Knowledge" by Elizabeth D. Liddy, Woojin Paik, Edmund Su, and Mary McKenna, RIAO '94 Proceddings.

"Intelligent text processing, and intelligent tradecraft," by Michael L. Weiner and Elizabeth D. Liddy, The Journal of AGSI, July 1995, also at http://www.mnis.net/agsi.html

Appendix: Comparative Table

Product Name →	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
Features						
Applications						
1.Customer Support	–	YES	–		YES	–
2.Decision Support	–	YES	–		YES	–
3.Helpline	–	YES	–		YES	–
4.Routing	YES		?	YES	YES (InRoute)	YES
5. Searching	YES		YES	YES	YES (InQuery)	YES
6.Training	–	YES	–		YES	
7. Classification	YES	YES	YES	YES	YES	YES (future)
8. Summarization	YES	–	YES	YES (LinguistX)	YES	YES (future)
Technology						
Adaptive Learning	YES	YES	?	YES	YES	SOME
AI	?	YES	SOME	YES	YES	SOME
CASE	NO	YES	NO	NO	?	NO
Expert Systems (rules)	NO	YES	YES	YES	NO	NO
Neural Network	YES	NO	NO	NO	NO	YES
NLP	NO	NO	YES	YES	NO	YES
Vector analysis	YES	NO	?	YES	YES	?

Levels of Linguistic Analysis	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
1. Morpho-logical	YES	YES	YES	YES	YES	YES
2. Lexical	YES	YES	YES	YES	YES	YES
3. Syntactic	NO	NO	YES	YES	NO	YES
4. Semantic	YES	NO	NO	YES	NO	YES
5. Discourse	NO	NO	NO	YES	NO	YES
6. Pragmatic	NO	NO	NO	YES	NO	YES
7. Multi-lingual	YES (any)	YES (14)	YES (11 lang.)	YES (LinguistX)	YES	YES

Linguistic Analysis Terminology

1. Morphological - the simplest parts (which can often have the most profound changes in meaning), e.g. literate vs. illiterate, sincere vs. insincere,

2. Lexical - the word forms themselves and how small changes in spelling cause drastic changes in meaning. Just think of Mrs. Malaprop.

3. Syntactic - the effect of word order on meaning, e.g. Iraq invades Kuwait, Kuwait invades Iraq

4. Semantic - the many different meanings that can be associated with a word or phrase, e.g. Clinton Beats Dole

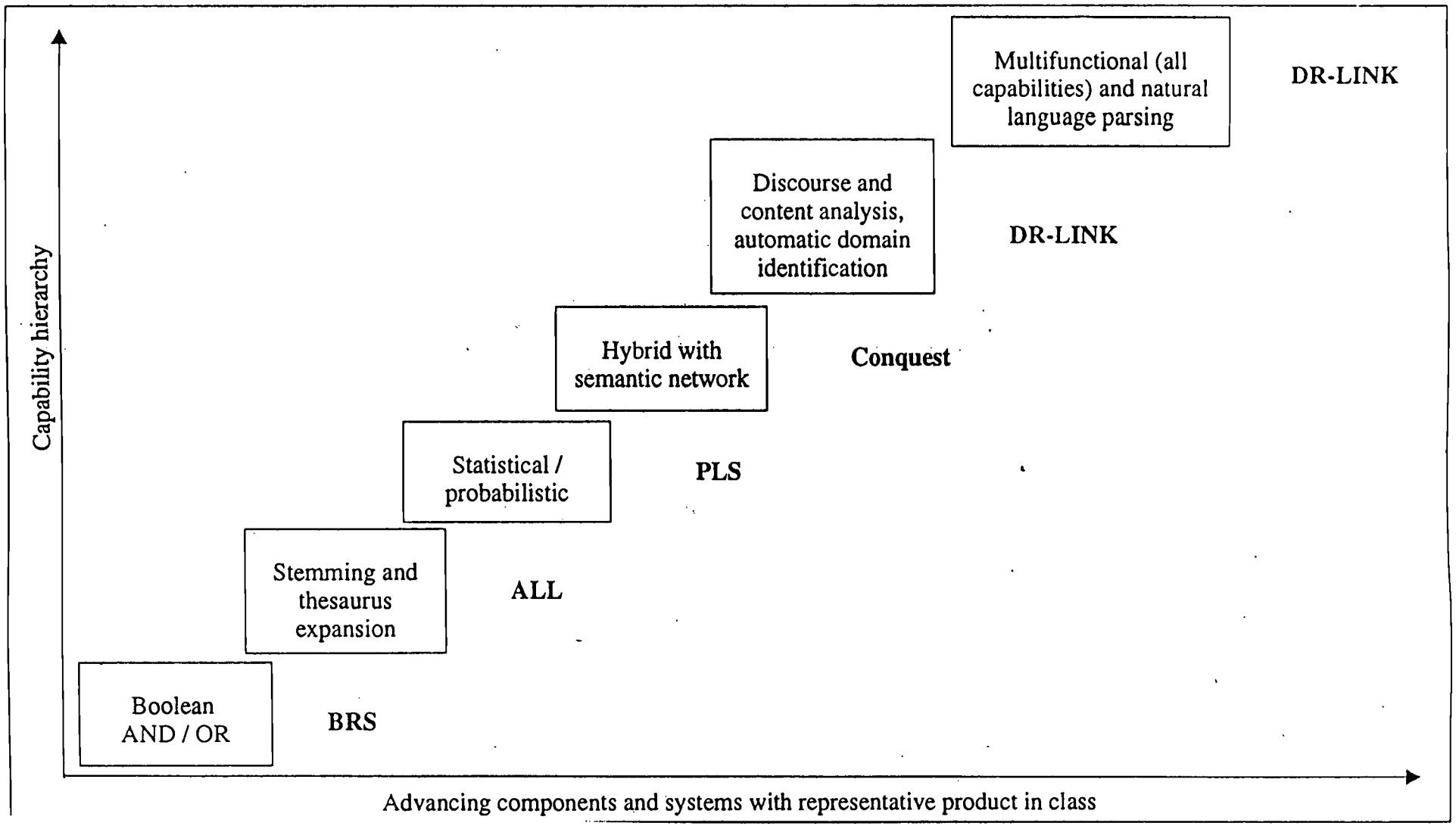
5. Discourse: all the meaning that is derived from the relationships between sentences, i.e., the cumulative meaning of a text

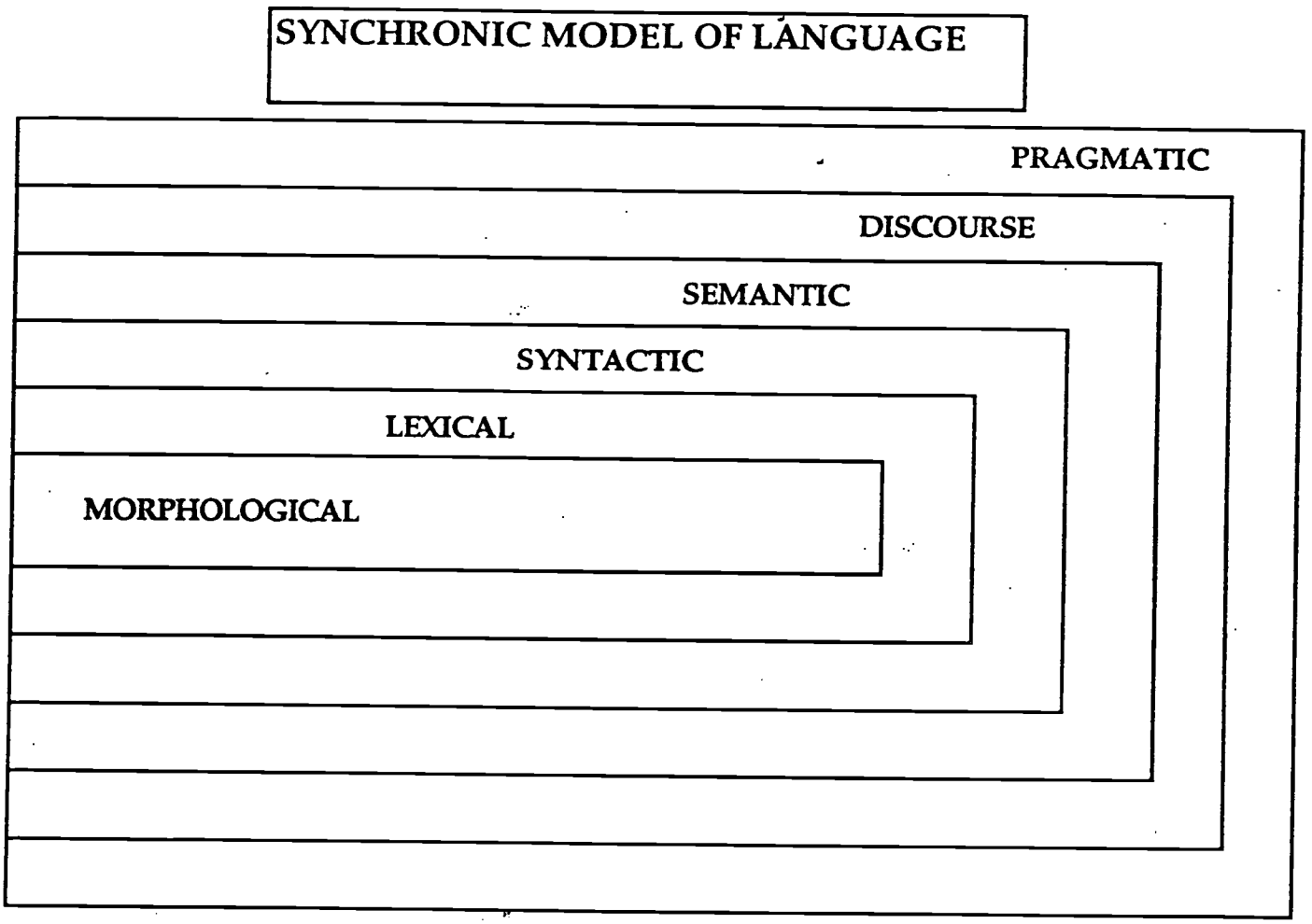
6. Pragmatic: that which is known outside the text and applied in understanding or interpreting the text, i.e., situational understanding. For example, when asked, "Do you know the time?," correct answer is "2:05," not "Yes, I do know the time."

7. Multilingual - operates with more than one language (not necessarily across languages).

Query Functionality	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
Boolean	YES	YES	?	YES	YES (InQuery)	YES
Proximity	YES	YES	?	YES	YES (InQuery)	NO
Stemming	YES	YES	YES	YES	YES (InQuery)	YES
Wildcards	?	YES		YES	NO	?
Fuzzy macthing	?	YES	YES	YES	NO	?
Phonetic	NO	?		YES	NO	?
Query expansion	YES	YES	YES			YES
Thesaurus	NO	NO	YES	YES	YES	YES
Phrase identification	YES	YES	YES	YES	NO	YES
Relevancy	YES	YES	?	YES	YES (InQuery)	YES
Assignable weights	NO	YES	?	YES	YES (InQuery)	NO
Semantic net	YES	NO	YES	YES	NO	YES
Discourse & content analysis	NO	YES	NO	SOME	NO	YES
Domain Identification	NO	YES	?	YES	NO	YES
Cross-domain	YES	?	YES	YES	YES	YES
NLP input	NO	YES	YES	YES	NO	YES
Cause/effect		YES				YES
Time Frames		YES	YES			YES

Search Features	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
1. Inter-active	YES	YES	YES	YES	NO	YES
2. Browsing	YES (cluster trees)	YES	YES			NO
3. By example	YES	YES	YES			YES
4. Disam-biguation by context	YES	?	?	YES	NO	YES
5. Geogra-phic name expansion	NO	?	?	YES	NO	YES
6. Visuali-zation	YES	?	YES			YES (future)
Administra-tion Features	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
APIs	?	?	YES	YES	YES	NO
System Features	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
Fully integrated	NO	NO	NO	YES	NO	YES
Relational	YES	YES	YES (ConText)	YES	YES	NO
Object Oriented	NO	NO	NO	NO	NO	?
Scalable	YES	?	YES	YES	YES	YES
No training required	NO	NO	YES	YES	YES	YES
Human intervention	OPTIONAL	YES	OPTIONAL	OPTIONAL	NO	YES
Customers	Convectis	CBR Content Navigator	LinguistX	ConText	InClass	DR-LINK
	DataTimes InfoSeek Intell.Xx Jostens Learning	AT&T Black & Decker Broderbund Canon Hewlett Packard PeopleSoft	InfoSeek Oracle SoftQuad Verity AOL	NetGuide Live http://www.netguide.com PR NewsWire at http://www.prnewswire.com/	InfoSeek West IBM/ Lotus	Patent Office


Capability hierarchy
Boolean AND / OR
BRS
Stemming and thesaurus expansion
ALL
Statistical / probabilistic
PLS
Hybrid with semantic network
Conquest
Discourse and content analysis, automatic domain identification
DR-LINK
Multifunctional (all capabilities) and natural language parsing
DR-LINK
Advancing components and systems with representative product in class


SYNCHRONIC MODEL OF LANGUAGE
PRAGMATIC
DISCOURSE
SEMANTIC
SYNTACTIC
LEXICAL
MORPHOLOGICAL

Library of Congress Experimental Search System

BASIC SEARCH | ADVANCED SEARCH | NUMBER SEARCH | BROWSE SEARCH | HELP

1 Enter SEARCH Type search words in the box below.

natural language processing [SEARCH] [CLEAR]

Search

Variants:	◉ Search exact words ○ Search word variants, e.g. plurals.
Languages:	◉ English only ○ All Languages
Fields:	☐ Titles ☐ Subjects ☐ Authors ☐ Notes (If no boxes are checked, all fields will be searched.)

2 SELECT Collections If no boxes are checked, all collections will be searched.

Cataloging Records	☐ Books (9,403,047)	☐ Maps (166,956)	☐ Serials (795,467)	☐ Prints and Photographs (68,135)
	☐ Manuscripts (updating)	☐ Music (203,294)	☐ Films (updating)	☐ Software (5,099)
Multimedia	☐ Legislation (173,528)	☐ Online Books (3,138)		☐ American Memory (76,701)

3 LIMIT Your Search (Optional)

Author, Date, Publisher Limits

Author

Publication Date

0000 through 1997

Publisher: Name/Location

Select Language(s) Limits

All
English
German
Spanish
French
Russian
Italian
Portuguese

To select more than one language, press and hold the **CTRL** key.

Select Category Limits

○ Fiction
○ Non-fiction
◉ Both

☐ Autobiography
☐ Biography
☐ Conferences
☐ Dictionaries
☐ Guidebooks
☐ Juvenile Literature

Go to Basic Search | [SEARCH] [CLEAR] | Go to Number Search

This is an **experimental** online public access catalog which is still under **development.**
It is updated daily

Comments:
Send comments about the **experimental search system** to: ess@loc.gov.
Send **general** questions about the library or questions of a research nature to: lcweb@loc.gov

21	Readings in **natural language processing** 1986 edited by Barbara J. Grosz, Karen Sparck Jones, Bonnie Lynn Webber. [P98 .R43 1986]
22	Processes, beliefs, and questions : essays on formal semantics of **natural language** and **natural language processing** 1982 edited by Stanley Peters and Esa Saarinen. [P325 .P7 1982]
23	Connectionist **natural language processing** : readings from Connection science 1992 edited by Noel Sharkey. [QA76.9.N38 C66 1992]
24	Reversible grammar in **natural language processing** 1994 edited by Tomek Strzalkowski. [QA76.9.N38 R48 1994]
25	**Natural language processing** and speech technology : results of the 3rd KONVENS Conference, Bielefeld, October 1996 1996 edited by Dafydd Gibbon. [P98 .K623 1996]
26	Evaluating **natural language processing** systems : an analysis and review 1995 Karen Sparck Jones, Julia R. Galliers. [QA76.9.N38 S74 1995]
27	Knowledge systems and Prolog : a logical approach to expert systems and **natural language processing** 1987 Adrian Walker (editor) ... [et al.]. [QA76.76.E95 K58 1987]
28	Subsymbolic **natural language processing** : an integrated model of scripts, lexicon, and memory 1993 Risto Miikkulainen. [QA76.87 .M54 1993]
29	Prolog for **natural language processing** 1991 Annie Gal ... [et al.]. [QA76.73.P76 P78 1991]
30	**Natural language processing** technologies in artificial intelligence : the science and industry perspective 1989 Klaus K. Obermeier. [Q336 .O24 1989]
31	**Natural language** engineering. 1995 [QA76.9.N38 N37]
32	TINLAP-2. theoretical issues in **natural language processing**-2, University of Illinois at Urbana-Champaign, July 25-27, 1978 1978 David L. Waltz, general chairman ; sponsored by the Association for Computational Linguistics, the Association for Computing Machinery, SIGART (ACM Special Interest Group in Artificial Intelligence). [P98 .T2]
33	**Natural language processing** : the PLNLP approach 1993 edited by Karen Jensen, George E. Heidorn, Stephen D. Richardson. [QA76.9.N38 N385 1993]
34	Computational models of **natural language processing** 1984 edited by Bruno G. Bara and Giovanni Guida. [P98 .C6123 1984]
35	From **natural language processing** to logic for expert systems : a logic based approach to artificial intelligence 1991 editor, André Thayse ; authors, Jean-Louis Binot ... [et. al.]. [QA76.9.N38 F76 1991]
36	Readings in automatic **language processing**. 1966 edited by David G. Hays. [P98 .H37]
37	**Natural language processing** 1994 edited by Fernando C.N. Pereira and Barbara J. Grosz. [QA76.9.N38 N384 1994]
38	Planning English sentences 1985 Douglas E. Appelt. [P98 .A67 1985]
39	Memory and context for **language** interpretation 1987 Hiyan Alshawi. [P98 .A52 1987]
40	Optimization of **natural** communication systems 1977 by Olga Akhmanova. [P91 .A42]

Item 31 of 194 *Serials* Browse the Shelf

Natural language engineering.

Cambridge, UK : Cambridge University Press, c1995-
v. : ill. ; 25 cm.
Four times a year
Vol. 1, pt. 1 (Mar. 1995)-
Title from cover.

Subjects:

Natural language processing (Computer science)--Periodicals.
Software engineering--Periodicals.

Language	Call Number	LCCN	Dewey Decimal	ISBN/ISSN
English (eng)	QA76.9.N38 N37	96658608	006.3/5	-

View the MARC Record for this item.

NEXT | PREVIOUS | ITEM LIST | BASIC SEARCH | ADVANCED SEARCH | NUMBER SEARCH | BROWSE SEARCH

The Library of Congress Experimental Search System

Item 37 of 194 Browse the Shelf

Natural language processing

edited by Fernando C.N. Pereira and Barbara J. Grosz.
Cambridge, Mass. : MIT Press, 1994.
vi, 531 p. ; 26 cm.
"A Bradford book."
"Reprinted from Artificial intelligence, an international journal, volume 63, numbers 1-2, 1993"--T.p. verso.
Includes bibliographical references and index.

Series:

Special issue of Artificial intelligence, an international journal

Subjects:

Natural language processing (Computer science)

Search for other works by:

Pereira, Fernando N. C.
Grosz, Barbara J.

Language	Call Number	LCCN	Dewey Decimal	ISBN/ISSN
English (eng)	QA76.9.N38 N384 1994	93039575	006.3/5	026266092X (acid-free paper)

View the MARC Record for this item.

NEXT | PREVIOUS | ITEM LIST | BASIC SEARCH | ADVANCED SEARCH | NUMBER SEARCH | BROWSE SEARCH

Categories (3)	
Category	**Score**
Web search tools	0.63
Communications & telecommunications	0.58
Computer science & engineering	0.56

Summaries (3)	
Summary	**Score**
DO WE STILL NEED CONTROLLED VOCABULARY?	0.63
Instead, a constructive merger of the best of both worlds - the fulltext analysis provided by web search engines and the controlled vocabularies found in library OPACs is recommended in this paper.	0.56
Thus, users will launch Web searches to find controlled vocabulary terms in OPACs and then use the controlled vocabulary to launch new Web searches.	0.59

Keywords (12)	
Keywords	**Score**
CBR	0.38
metadata	0.36
fulltext	0.31
Inference	0.30
Convectis	0.30
LinguistX	2.00
InClass	2.00
LIVT	4.00
webtime	3.00
LCSH	3.00
Aptex	5.00
HNC	2.00

Convectis Document Summary (minus tables)

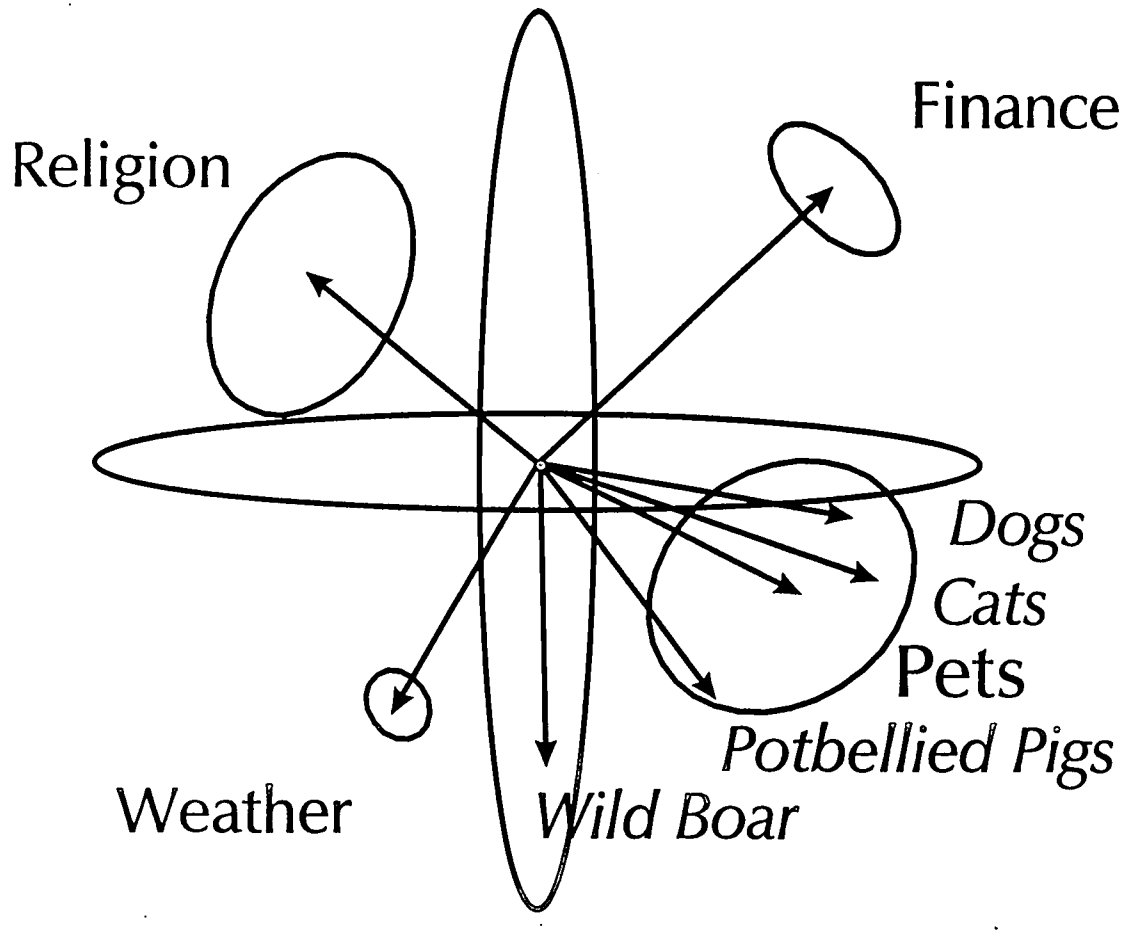
Categories (3)	
Category	**Score**
Web search tools	0.43
Searching & exploring the Internet	0.38
Bioinformatics	0.37

Summaries (3)	
Summary	**Score**
Several commercial software products, many affiliated with major search engine vendors, claim to have moved beyond fulltext retrieval based on simple word-matching to more sophisticated techniques capable of supporting automatic classification and analysis of fulltext documents equal or superior to that provided by human indexers and catalogers.	0.42
With respect to fulltext documents, can these tools increase cataloger productivity by presenting controlled vocabulary terms for de-selection and by refocusing the cataloger's energies on the editing of machine-generated records and the maintenance of software programs which generate such records?	0.44
Several commercial products having potential for improved subject access to fulltext and for automatically or semi-automatically "cataloging" fulltext are reviewed within the context of other existing strategies for indexing and organizing materials on the Internet.	0.40

Keywords (5)	
Keywords	**Score**
fulltext	0.60
cataloger	0.38
vocabulary	0.38
traditional	0.21
libraries	0.20

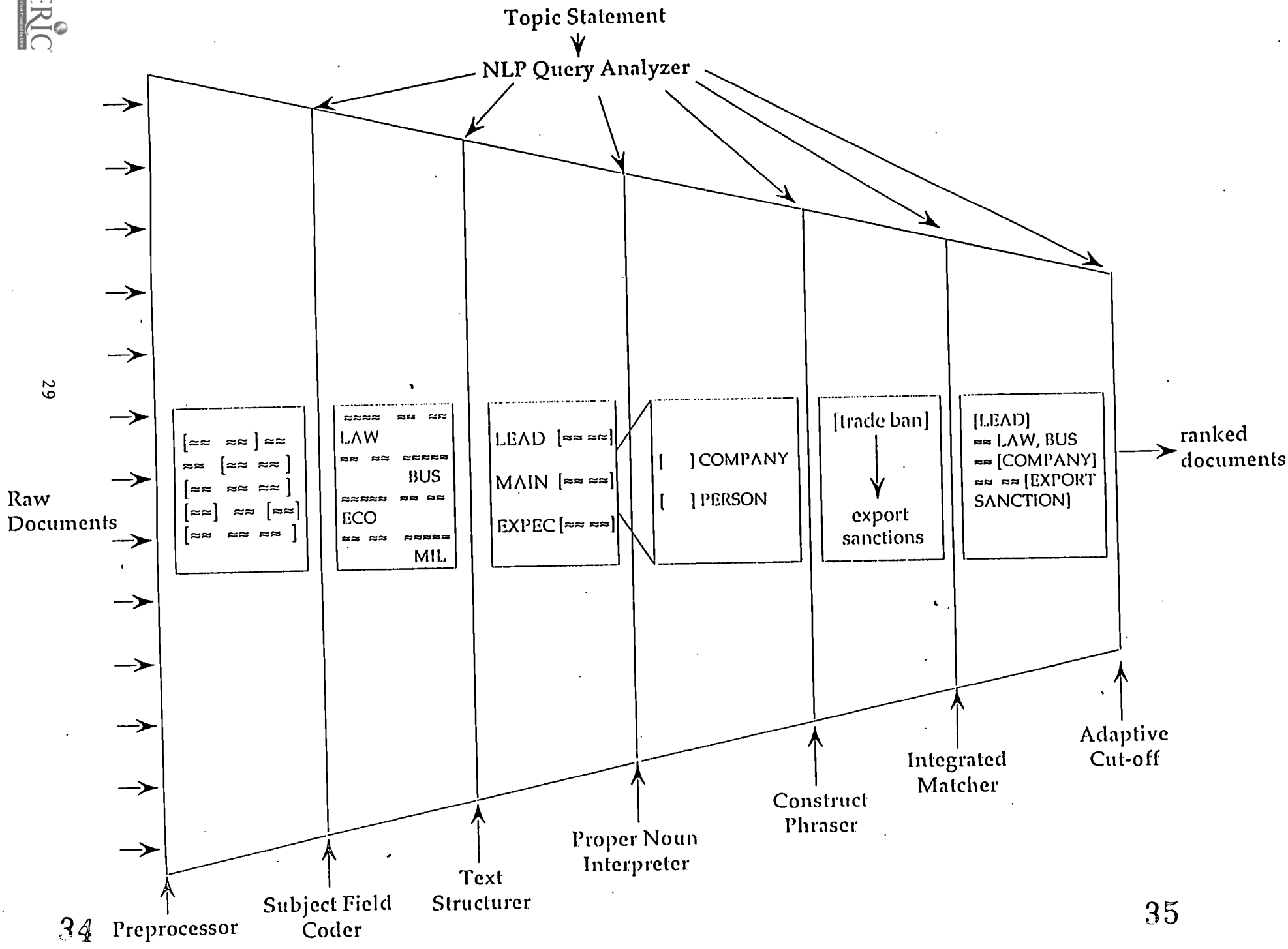
Convectis Abstract Summary

Context Vector Representation



- Words with similar meaning have Context Vectors that point in similar directions
- Documents on similar topics also have Context Vectors pointing in similar directions
- The distance between Context Vectors equals the degree of similarity

APTEX
AN HNC COMPANY



DR-LINK

Figure 1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").