TOAS Intelligence Mining; Analysis of Natural Language Processing and Computational Linguistics

Robert J. Watts[1], Alan L. Porter, Ph.D., Scott Cunningham, and Donghua Zhu[2]

[1] Tank-automotive and Armaments Command, National Automotive Center, Warren, Michigan 48397-5000, USA
[2]Technology Policy and Assessment Center, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

Abstract: The Technology Opportunities Analysis System (TOAS), being developed under a Defense Advanced Research Projects Agency (DARPA) project, enables mining of text files using bibliometrics. TOAS, a software system, extracts useful information from literature abstract files, which have identified fields that repeat in each abstract record of specific databases, such as Engineering Index (ENGI), INSPEC, Business Index, U.S. Patents, and the National Technical Information Service (NTIS) Research Reports. The TOAS applies various technologies, which include natural language processing (NLP), computational linguistics (CL), fuzzy analysis, latent semantic indexing, and principle components analysis (PCA). This software system combines simple operations (i.e., listing, counting, list comparisons and sorting of search term retrieved consolidated records' field results) with complex matrix manipulations, statistical inference and artificial intelligence approaches to reveal patterns and provide insights from large amounts of information, primarily related to technology-oriented management issues.

The authors apply the TOAS tool on its own root technologies, NLP and computational linguistics--two apparently synonymous terms. These terms, however, when used in a literature search of the same abstract databases, ENGI and INSPEC, provide distinctly different search results with only 10% to 25% search result abstract records overlap. This paper introduces TOAS, summarizes analyses comparing NLP and CL, and then discusses the underlying development implications.

1 Introduction: TOA™ has been under development at Georgia Tech since 1990 (Porter and Detampel, 1995). TOA presumes that useful evidence on the prospects for technological innovation can be gleaned from bibliometric analyses (Watts and Porter, under submission). For brevity, as well as clarity, TOAS familiarization will be accomplished with the comparative bibliometric analysis of "natural language processing" and "computational linguistics", two technologies critical to the development of the TOAS software. TOAS is a monitoring and bibliometric tool kit that develops intelligence on a chosen topic by establishing and assessing publication and/or patenting patterns. The TOAS does not function as a search engine; but receives files containing the records assembled during a database search. The TOAS can be tailored to analyze fixed field text database records by predefining the database field descriptors and delineators. Once the field formatting is accomplished, the TOAS retains this record analysis interface information for future search result analyses.

2 NLP / CL Analysis: Our quest to compare NLP and CL begins with a Dialog search comparison of several promising literature abstract databases. Table 1 summarizes the

Table 1 - Database records containing "Natural Language Processing" or "Computational Linguistics"

Database	NLP	CL
INSPEC (1983-97)	1626	4386
EI Compendex (1970-97)	1411	4067
Linguistics & Language Behav. Abs (1973-96)	529	3305
Info Sci Abs (1966-97)	293	119
IAC Computer Database (1983-97)	306	60
Dissertation Abs (1861-97)	173	59
Library & Info Sci (1969-97)	183	56
SciSearch (1988-97)	467	41
Social SciSearch (1972-97)	128	35
Microcomputer Abs (1974-97)	43	3

prevalence of records (abstracts) on NLP and computational linguistics (CL). Based on this, we pursue INSPEC as our primary resource for bibliometric comparison of the two technology disciplines.

The TOA process combines analyses of the literature abstract database field information with expert review and synthesis (Porter and Detampel, 1995). The process, which is iterative in nature, includes domain specification, data acquisition and analysis. The domain specification, which will be called the technology space for purpose of this paper, depicts a most important and critical step in the process. In defining the technology space, the researcher adds context to the search term and gains an awareness of the differences between the search term file records' summaries and the referenced scientific fields or areas of interest.

The search term defines the sample population on which inferences will be made on the field of interest. For example, the search on NLP yielded 1398 abstract records; that for CL, 3867. [Note: analyses conducted on results of simple searches, "natural adjacent to language adjacent to processing" and "computational adjacent to linguistics".] Two abstract files, one each for the NLP and CL records, have been created and analyzed using the TOAS software. Table 2 summarizes the most frequent subject index terms (keywords") for each file. This summary presentation aids recognition of principle topics within the subject matter addressed. Asterisks indicate the terms common to NLP and CL file records. Note that all 64 of the most frequently used keywords from each file, NLP and CL, are common to the other file. Further, of the 1398 NLP records, 290 also contain the term CL, 129 contain the phrase "knowledge based systems", 127 "artificial intelligence" and 29 "learning systems". These highlighted terms from Table 2 have been used as separate and combined search constructs in the ENGI and INSP databases. Table 3 presents the synopsized search results. The summary from Table 2 indicates that artificial intelligence may be a sub-element of both NLP and CL. Table 3 unveils the breadth of the AI field documentation. Observe in Table 3 that 323 INSP records contain both search terms, NLP and CL; while 1398 contain NLP and 3867 include CL. The summary on the right side of Table 3 provides the search result overlap calculation for the various terms; note that only 23.1% of the records retrieved with the term NLP, also contain the term CL. Likewise, 8.4% assembled using the phrase CL contain the NLP term. This low level of co-occurrence of phrases, yet strong commonality of keywords used in the two files for NLP and CL, as depicted in Table 2, create an odd dichotomy. A

Table 2 - TOAS Field Summary Lists and List Comparison Features

SEARCH TERM KEYWORD LISTS and COMPARISONS					
Count		Natural Language Processing Abstracts' Keywords	Count		Computational Linguistics Abstracts' Keywords
944	*	NATURAL LANGUAGES	3526	*	COMPUTATIONAL LINGUISTICS
290	*	COMPUTATIONAL LINGUISTICS	1461	*	NATURAL LANGUAGES
285	*	GRAMMARS	821	*	GRAMMARS
147	*	NATURAL	341	*	LANGUAGE TRANSLATION
145	*	LINGUISTICS	296	*	LINGUISTICS
129	*	KNOWLEDGE BASED SYSTEMS	295	*	LANGUAGES
127	*	ARTIFICIAL INTELLIGENCE	219	*	KNOWLEDGE REPRESENTATION
124	*	USER INTERFACES	213	*	NATURAL
125	*	LANGUAGES	193	*	FORMAL LOGIC
112	*	EXPERT SYSTEMS	168	*	LOGIC PROGRAMMING
104	*	LANGUAGE TRANSLATION	164	*	FORMAL LANGUAGES
102	*	KNOWLEDGE REPRESENTATION	164	*	INFERENCE MECHANISMS
95	*	NATURAL LANGUAGE INTERFACES	146	*	KNOWLEDGE BASED SYSTEMS
87	*	INFERENCE MECHANISMS	127	*	PROGRAMMING
88	*	INFORMATION RETRIEVAL	118	*	COMPUTATIONAL COMPLEXITY
80	*	NEURAL NETS	108	*	USER INTERFACES
65	*	KNOWLEDGE ENGINEERING	107	*	PROGRAMMING THEORY
63	*	DATABASE MANAGEMENT SYSTEMS	107	*	FUZZY SET THEORY
63	*	LOGIC PROGRAMMING	103	*	ARTIFICIAL INTELLIGENCE
60	*	SPEECH RECOGNITION	104	*	SPEECH RECOGNITION
58	*	GLOSSARIES	90	*	WORD PROCESSING
54	*	KNOWLEDGE ACQUISITION	89	*	FUZZY LOGIC
46	*	WORD PROCESSING	89	*	LOGIC
42	*	INFORMATION RETRIEVAL SYSTEMS	85	*	DATA STRUCTURES
41	*	INDEXING	84	*	FORMAL SPECIFICATION
41	*	FORMAL LOGIC	83	*	CONTEXT-FREE GRAMMARS
40	*	KNOWLEDGE	75	*	THEORY
37	*	COMPUTER AIDED INSTRUCTION	76	*	KNOWLEDGE
37	*	SYSTEMS	72	*	PARALLEL PROGRAMMING
37	*	PROCESSING	71	*	KNOWLEDGE ENGINEERING
34	*	NATURAL LANGUAGE	71	*	FORMAL
34	*	INTERFACES	70	*	EXPERT SYSTEMS
32	*	LEARNING (ARTIFICIAL INTELLIGENCE	69	*	SYSTEMS
29	*	LEARNING SYSTEMS	68	*	HIGH LEVEL LANGUAGES
26	*	LANGUAGE INTERFACES	65	*	PROLOG
25	*	PROLOG	64	*	NATURAL LANGUAGE INTERFACES
25	*	DEDUCTIVE DATABASES	64	*	PROCESSING
24	*	HYPERMEDIA	63	*	NEURAL NETS
23	*	REPRESENTATION	63	*	SPECIFICATION LANGUAGES
23	*	COMPUTER VISION	61	*	OBJECT-ORIENTED PROGRAMMING
22	*	SPEECH ANALYSIS AND PROCESSING	60	*	KNOWLEDGE ACQUISITION
21	*	PROBABILITY	59	*	DATABASE THEORY
21	*	DATA STRUCTURES	59	*	TREES (MATHEMATICS
20	*	FORMAL LANGUAGES	56	*	PROBABILITY
22	*	SPEECH	56	*	GLOSSARIES
20	*	STATISTICAL ANALYSIS	56	*	COMPUTATIONAL
20	*	INFORMATION ANALYSIS	55	*	THEOREM PROVING
19	*	DECISION SUPPORT SYSTEMS	53	*	PROGRAM COMPILERS
19	*	CONTEXT-FREE GRAMMARS	53	*	SET THEORY
19	*	PARALLEL PROCESSING	51	*	SPEECH
19	*	CLASSIFICATION	49	*	PARALLEL LANGUAGES
19	*	CONSTRAINT HANDLING	49	*	INFORMATION RETRIEVAL
18	*	INTERACTIVE SYSTEMS	49	*	DATABASE MANAGEMENT SYSTEMS
18	*	PROGRAMMING	48	*	RELATIONAL DATABASES
18	*	RESEARCH INITIATIVES	47	*	PROCESS ALGEBRA
17	*	COGNITIVE SYSTEMS	47	*	CONSTRAINT HANDLING
17	*	COMPUTATIONAL COMPLEXITY	47	*	DEDUCTIVE DATABASES
17	*	SEMANTIC NETWORKS	46	*	PARALLEL
17	*	MEDICAL COMPUTING	44	*	INTERACTIVE SYSTEMS
16	*	QUERY LANGUAGES	44	*	INFERENCE
16	*	FORMAL SPECIFICATION	43	*	TYPE THEORY
16	*	QUERY PROCESSING	43	*	LITERATURE
16	*	SPEECH SYNTHESIS	43	*	MECHANISMS

Table 3 - Search Terms' Abstracts and Count Comparisons

Term #	Search Term	Number of Abstracts ENGI	INSP
1	Artificial adj Intelligence	25,881	39,701
2	Knowledge adj Based adj System	9,488	18,295
3	Learning adj System	10,170	8,989
4	**Natural adj Language adj Processing**	1,306	**1,396**
5	**Computational adj Linguistics**	3,864	**3,867**
6	1 and 2	3,989	5,160
7	1 and 3	1,722	2,177
8	1 and 4	557	448
9	1 and 5	505	661
10	2 and 3	801	676
11	2 and 4	197	174
12	2 and 5	385	213
13	3 and 4	106	38
14	3 and 5	157	38
15	4 and 5	309	**323**

ENGI (Column Intersect Qty with Row Pub Qty)	Artificial adj Intelligence	Knowledge adj Based adj System	Learning adj System	Natural adj Language adj Processing	Computational adj Linguistics
Artificial adj Intelligence	1.000	0.154	0.067	0.022	0.020
Knowledge adj Based adj System	0.420	1.000	0.084	0.021	0.041
Learning adj System	0.169	0.079	1.000	0.010	0.015
Natural adj Language adj Processing	0.426	0.151	0.081	1.000	0.237
Computational adj Linguistics	0.131	0.100	0.041	0.080	1.000

INSP (Column Intersect Qty with Row Pub Qty)	Artificial adj Intelligence	Knowledge adj Based adj System	Learning adj System	Natural adj Language adj Processing	Computational adj Linguistics
Artificial adj Intelligence	1.000	0.130	0.055	0.011	0.017
Knowledge adj Based adj System	0.282	1.000	0.037	0.010	0.012
Learning adj System	0.242	0.075	1.000	0.004	0.004
Natural adj Language adj Processing	**0.320**	**0.124**	**0.027**	**1.000**	**0.231**
Computational adj Linguistics	0.171	0.055	0.010	**0.084**	1.000

dichotomy of terms that drives our analyses.

The TOAS file records' field summary listings and comparative analyses provide preliminary insights on social, intellectual and temporal indicators of research (Cunningham, 1996). Astute analyses reveal hierarchical relationships within the system of interest, functionality and engineering principles performed, dependency relationships with other technologies and functions, applications areas, related materials and enabling factors within the development environment. The next section uses the field summary listings to explore the nature of NLP and CL activities.

2.1 Field Summary Listing Operations: This section explores the question of whether NLP research differs from CL research. The search on NLP yielded 1398 items (1986-1996); that for CL, 3867. Of those, 323 items contained both terms -- NLP and CL. For cleanest comparison, four data sets are created:

A. NLP (not CL) dated 1995 or 1996 -- 170 abstracts
B. NLP (not CL) dated 1986, 1987, or 1988 – 204 abstracts
C. CL (not NLP) dated 1995-1995 – 456 abstracts
D. CL (not NLP) dated 1986, 1987, or 1988 – 562 abstracts

We examine these four data sets first in terms of commonality of file field summary listings: subject index terms (keywords), authors, author affiliations, and sources (journals, proceedings) (Kostoff, 1993). How similar are they? Comparing sets A and C, we find that all 34 of the most common keywords in set A (NLP 1995-96) also appear in set C (CL 1995-96)! Indeed, only 11 of the 111 most common NLP keywords do not appear in the CL items. Only one of those appears 5 or more times in the NLP set, but not at all in the CL set – "learning by example". This suggests strong commonality.

Conversely, while a majority of the most frequent CL keywords also appear in the NLP items (71 of 113), many do not (42 of 113). This stirs us to probe further. We list the most frequent NLP keywords in the left column of a 2-column table with CL keywords in the right column (similar to Table 2). We then link the most frequent 27 NLP keywords to the same word in the CL listing. A few of the most basic terms head both lists: natural languages, natural, grammars, and languages. Three of the other NLP top 27 appear higher ranked in the CL list: language translation, linguistics, and formal logic. The remainder of the NLP top 27 rank considerably lower in the CL listing. Conversely, most all of the other top 32 in the CL list appear relatively infrequently or not at all in the NLP list. This suggests that there are apparently different emphasis areas in these two sets of abstracts, even though they share so much common ground.

We next compare where the respective NLP and CL abstracts were published. Comparing the most frequent 1995-96 outlets for NLP (>1 publication) and CL (>3, because the sample is considerably larger), shows only modest overlap. Of the 20 sources of more than one NLP article, only one is a heavy CL source (Trans. of the Information Processing Society of Japan); another 7 do have between 1 and 3 items using the term CL. Conversely, of the 26 sources publishing more than three CL articles, only one is a heavy NLP outlet, as noted, and only 3 others have published a single item containing the phrase NLP. This suggests that NLP and CL are not smoothly intermingled; rather, they seem to act as two distinct communities with overlapping interests.

Who does research in the two sets? Comparing the affiliations of the authors finds

that most of the CL institutions link to computer science. Of the 37 institutions publishing at least 2 papers, 1995-96, 21 explicitly include "Computer Science" in their title. The others include: information science, information, computing, linguistics, information & operations research, systems science, industrial engineering, and art & design, plus Bell Labs. Some 34 of the 37 are clearly academic. On the other side, the 12 NLP institutions include Bell Labs, 3 computer science departments, and a smattering of information and/or technology-oriented others. More than half are non-academic.

Of the 12 leading NLP institutions and 37 leading CL producers in 1995-96, only two are the same: Bell Labs and the University of Pennsylvania's Department of Computing and Information Science. Two more of the 12 NLP institutions also have CL publications. Only 1 of the remaining 35 CL producers shows as also producing an NLP paper.

Individuals who publish using the phrase NLP are not likely to publish using CL. (Recall that the samples being compared eliminate the articles containing both phrases.) We focus on the 24 CL authors publishing at least 3 items and the 28 NLP authors publishing at least 2. Of these, only 3 are common to both lists. Of the other 25 NLP authors, only 4 have another 1995-96 publication using the phrase CL. Of the other 21 CL authors, none have another publication using NLP.

Taken together, examination of summary listings of keywords, sources, institutional affiliations, and authors suggests that there is great overlap in the subject matter of NLP and CL, yet these appear to derive from strikingly distinct scholarly communities.

3 Field Summary Lists' Comparisons and Co-occurrence Matrices: Inference analyses, as above, can be easily enabled by the TOAS summary lists, particularly when performed on a distinct segment of the data set (e.g., time slice or content limited). The TOAS expands the analysis capability by compiling and comparing field summary lists in matrices to reveal co-occurrences of terms and patterns within data. The list comparison matrices do add complexity to the interpretation process. However, matrix operations on complex tables, such as on co-word/document matrices, can uncover underlying structure and linkages within the data set analyzed. Singular value decomposition (SVD) represents a useful tool for data reduction of this type (Press and Flannery et al 1986) SVD underlies many forms of multivariate analysis, such as principle component analysis (PCA), factor analysis and correspondence analysis. The use of SVD in analysis of word and document occurrences has been incorporated into the technique, latent semantic indexing (Deerwester et al 1990). The output of SVD approximates the original data and, more importantly, reveals structural patterns within the data. Deerwester et al. (1990) apply SVD to relate words and documents in a bibliometric data set. Cunningham (1996) discusses its use in the classification of documents in large science policy databases.

Publication data often vary over orders of magnitude (e.g., term occurrence frequencies). High magnitude variables create greater potential for error introduction in the SVD least squares analysis. This error introduction can be reduced by normalizing the data. Pearson correlation is the normalization technique applied to the data by the TOAS software when performing the principle components analysis.

3.1 Principle Component Analysis of NLP and CL: To continue the bibliometric evaluation of the similarities and differences between NLP and CL, the top occurring keywords for the NLP and CL abstracts, Table 2, were input separately into the PCA

Table 4 - Top Level Keyword Groupings for NLP and CL

Computational Linguistics' Keywords Groupings	Natural Language Processing's Keywords Groupings
a. **language translation** - which includes "language translation" grouped with "grammars".	a. **CL** - "computational linguistics"
b. **natural language** - which includes "natural" and "languages".	b. **knowledge** - which includes "knowledge representation" and "inference mechanisms".
c. **fuzzy** - which includes "fuzzy set theory" and "fuzzy logic".	c. **linguistics** - which includes "glossaries" with "linguistics".
d. **program** - which includes "logic programming", "programming theory" & "programming".	d. **database management** - "database management systems".
e. **knowledge** - which includes "knowledge representation", "inference mechanisms" & "knowledge based systems".	e. **Logic**- which includes "knowledge engineering", "knowledge based systems" & "logic programming".
f. **formal** - which includes "formal languages" and "formal logic"	f. **Information** - "information retrieval".
g. **speech** - which is "speech recognition".	g. **expert/AI** - which includes "artificial intelligence" & "expert systems".
	h. **Acquire** - "knowledge acquisition"

routine (e.g., one PCA analysis was performed for the 1380 NLP abstracts and one for the 3830 CL abstracts). The number of factors extracted in the SVD was incremented until 50 percent of the term variation was explained by the analysis generated factors. The resulting factor group listings were then sorted using the keywords' associated term loading coefficients. High loading terms for each factor were then grouped together. This process identified the eight keyword groupings for NLP and seven groupings for CL as shown in Table 4.

The TOAS allows the user to create abstract files which represent subsets of an initially processed file of abstracts; the subset being termed a transitional file. The transitional file contains all abstracts that include the user designated summary list terms (i.e., referring to Table 2, the NLP transitional file for abstracts containing the term "computational linguistics" includes 290 abstracts). Transitional files, once created, can be processed and the subset abstract field summary list operations can be performed in the same manner as on the master file. Transitional files have been generated for each of the factor groupings identified in Table 4. TOAS aids in developing temporal summaries on research. Table 5, for example, provides the chronological publication histograms for the Table 4 keyword groupings' transitional files for CL, using actual publication dates. These tables provide relative rates of growth for the various sub-group areas of CL. Note in Table 5 that the number of publications that contained the sub-group terms for "language translation" and "speech recognition" peaked in 1989 and have all but disappeared in 1996. The emergence of the "natural language" sub-group publications in the 90's is also noteworthy. Observations such as these help to inspire hypotheses, which must then be followed up with further analyses and research.

The PCA analysis yields groupings of terms tending to occur together in abstract records (Table 4). These groupings can be used to create categories of abstract records. Field comparisons among the defined abstract groupings can indicate the degree of overlap (i.e., duplicate records). Tables 6 and 7 summarize the results of a title comparison analysis. Note that abstracts on conferences, symposium and "like" proceedings would use a broad range of keywords and create apparent overlap of

Table 5 - Publication Chronology - Computational Linguistics

Year Published	# Pubs CL *Language* Translation Group	% Total Language Translation Group	# Pubs CL Natural Language Group	% Total Natural Language Group	# Pubs CL *Fuzzy* Group	% Total *Fuzzy* Group	# Pubs CL *Programmin* g Group	% Total *Programmin* g Group	# Pubs CL *Knowledge* Group	% Total *Knowledge* Group	# Pubs CL Formal Languages & Logic Group	% Total Formal Languages & Logic Group	# Pubs CL Speech Group	% Total Speech Group
1985	4	0%	4	1%		0%		0%		0%		0%		0%
1986	106	10%	40	13%	2	1%	3	1%	1	0%	14	4%	1	1%
1987	59	6%	7	2%	8	5%	3	1%	4	1%	11	3%	1	1%
1988	63	6%	2	1%	14	8%	29	8%	29	6%	27	8%	5	5%
1989	190	18%		0%	11	7%	19	5%	52	11%	41	12%	26	25%
1990	111	10%	2	1%	12	7%	37	11%	56	12%	35	11%	8	8%
1991	139	13%	7	2%	11	7%	32	9%	60	13%	40	12%	10	10%
1992	84	8%	26	8%	18	11%	15	4%	66	14%	26	8%	10	10%
1993	71	7%	33	10%	21	13%	29	8%	32	7%	20	6%	13	13%
1994	99	9%	82	26%	21	13%	66	19%	57	13%	47	14%	22	21%
1995	74	7%	66	21%	28	17%	61	17%	45	10%	33	10%	7	7%
1996	58	5%	49	15%	19	11%	55	16%	34	7%	37	11%		0%
Totals	1058	100%	320	100%	167	100%	349	100%	456	100%	331	100%	103	100%

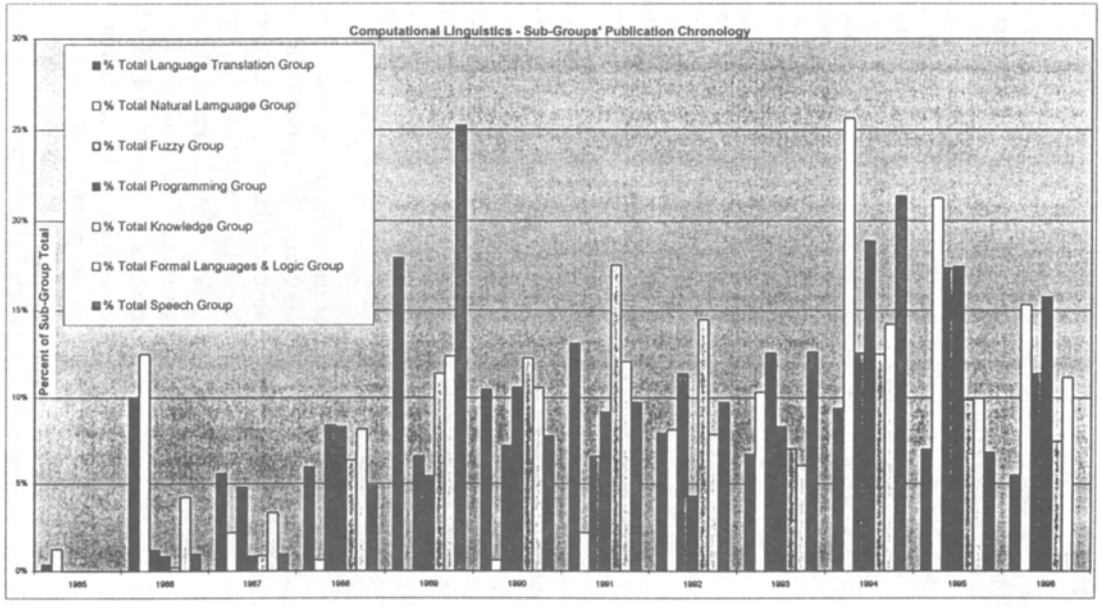

Table 6 - Computational Linguistics Sub-Group Comparisons

Row Overlap Percentage with Column Sub-Group

	Language Translation	Natural Language	Program	Fuzzy	Knowledge	Formal	Speech
Language Translation		8.3%	4.1%	0.0%	8.6%	4.9%	3.4%
Natural Language	27.5%		7.8%	1.9%	8.4%	6.6%	1.6%
Program	12.3%	7.2%		1.1%	7.7%	15.2%	0.0%
Fuzzy	1.8%	3.6%	2.4%		24.6%	3.6%	1.8%
Knowledge	20.0%	5.9%	5.9%	9.0%		9.9%	1.0%
Formal	15.7%	6.3%	16.0%	1.8%	13.6%		0.0%
Speech	35.0%	4.9%	1.0%	2.9%	3.9%	0.0%	

Table 7 - Natural Language Processing Sub-Group Comparisons

Row Overlap Percentage with Column Sub-Group

	Computational Linguistics	Knowledge	Linguistics	Database Management	Logic	Information Retrieval	Expert/AI	Knowledge Acquisition
Computational Linguistics		11.0%	10.3%	2.1%	12.1%	2.1%	4.1%	2.4%
Knowledge	18.8%		9.4%	1.8%	23.5%	2.4%	15.9%	7.6%
Linguistics	16.0%	8.6%		4.3%	11.2%	6.4%	18.4%	5.3%
Database Management	9.5%	4.8%	12.7%		22.2%	9.5%	19.0%	0.0%
Logic	15.2%	17.3%	9.1%	6.1%		5.6%	18.6%	6.1%
Information Retrieval	7.0%	4.7%	14.0%	7.0%	15.1%		14.0%	4.7%
Expert/AI	5.7%	12.9%	7.7%	5.7%	20.6%	5.7%		4.8%
Knowledge Acquisition	13.0%	24.0%	18.5%	0.0%	25.9%	7.4%	18.5%	

developed sub-group categories. In spite of embedded database noise/bias effects, the TOAS PCA process creates relatively discrete literature abstract sub-groupings -- a majority of the overlap percentages being under 10 percent. Considering that all the sub-group files were generated by the same top level search term, either "natural language processing" or "computational linguistics," the low level of overlap between sub-grouping file records is impressive. These sub-groupings help categorize the prevalent areas of interest within the documented research.

What can be inferred from this top level PCA analysis? Referring to Table 4, both documentation databases have groupings which refer to knowledge representation and inference mechanisms. The NLP literature includes application categories of information retrieval and knowledge acquisition. The CL application groupings include language translation and speech recognition. In regards to computer concerns, NLP addresses database management systems, knowledge based systems, logic programming and glossaries. CL appears to go deeper into the machine by including programming theory and programming. Computational linguistics literature refers more to fuzzy set theory/fuzzy logic, formal languages and formal logic; whereas NLP literature sites artificial intelligence and expert systems. One comparison exists, which raised our curiosity; NLP documentation breaks out computational linguistics as a sub-group, CL literature decomposes only to a natural language category. Do the CL natural language articles refer to NLP , computer languages, or something else?

To answer the last question, we compared two transitional files -- NLP-CL (290 abstract records) and CL-NL (320 abstracts). In comparing "title" lists, there are only 33 abstracts with common titles (e.g., about 10%). In comparing "keyword" lists, one finds 133 common of the 346 keywords of the CL-NL abstracts and the 189 keywords of the NLP-CL abstract grouping. Do we have two related fields performing R&D in isolation of one another? There are 44 common affiliations within the 206 organizations that published in the NLP-CL category and 259 CL-NL publication organizations. Table 8 lists the PCA high factor term groupings determined as described earlier, here using the two transitional files NLP-CL and CL-NL. Assessment comments have been noted at the bottom of the Table. Our top level PCA factor interpretations appear to be corroborated by this second transitional file PCA factor analysis.

Further investigations reveal that the transitional files created by marking the keyword terms as grouped in Table 8, segregate the records into average size abstract groups of twenty abstracts for NLP-CL categories and eleven abstracts for CL-NL sub-groups; certainly a manageable quantity for any researcher. Whether the abstract groupings are clearly related and define a distinct class of research/subject matter has yet to be determined. The questions raised during this research effort and in this paper will be answered. They must, for NLP and CL form the foundation of the TOAS system.

4 Conclusions: Similar analyses, as presented above for NLP-CL and CL-NL, are being performed on the other transitional files as defined by the Table 4 keyword groupings. Through these analyses we hope to further define the TOAS analysis software requirements. Likewise, we are also investigating the content of the NLP and CL files' abstracts which have not been captured by the Table 4 term groupings. The subject matter that gets omitted by the currently defined PCA factor groupings will be as important to the design updates of the applied algorithms as what gets included. Much research and associated software development remains ahead. The NLP and CL comparison analyses

Table 8 - NLP-CL and CL-NL Literature Sub-groups PCA Factor Groupings

NLP-CL Groupings

a. "Logic programming" and "PROLOG"
b. "Linguistics" and "Information Retrieval"
c. "Natural Language" and "Interfaces"
d. "Knowledge" and "Representation"
e. "Knowledge Representation", "Inference Mechanisms" and "User Interfaces"
f. "Semantic Networks", "Constraint Handling" and "Artificial Intelligence"
g. "Database Management Systems", "Statistical Analysis" and "Glossaries"
h. "Language Translation"

CL-NL Groupings

a. "Knowledge Representation"
b. "Object-Oriented Databases" and "Databases"
c. "Formal Specification", "Specification Languages", "Software Tools", "Compiler Generators" and "Formal"
d. "Language Interfaces"
e. "Natural Language Interfaces" and "Knowledge"
f. "Functional Programming" and "Functional"
g. "Rewriting Systems" and "Theorem Proving"
h. "Object-Oriented Languages" and "Parallel"
i. "Computational", "Abstract Data Types", "Object-Oriented Programming"
j. "Speech Recognition"
k. "Data"
l. "Attribute Grammars" and "High Level"
m. "User Interfaces" and "Expert Systems"
n. "Relational Databases" and "Database Theory"
o. "Lambda Calculus" and "Type Theory"

Note Similarities: Natural Language Interfaces, Knowledge Representation, User Interfaces

Note Differences:
1. NLP has Language Translation and CL has speech recognition and rewriting systems.
2. NLP applies linguistics, information retrieval, glossaries, semantic networks, constraint handling and artificial intelligence, whereas CL denotes attribute grammars, computational, abstract data types, and expert systems.
3. NLP references logic programming and PROLOG, while CL calls out object-oriented programming, object-oriented languages, functional programming, formal specification, specification languages, software tools, compiler generators.
4. NLP lists statistical analysis; CL sites Lambda calculus, type theory and theorem proving.
5. NLP has database management systems, while CL has object-oriented databases, relational databases, database theory.

have provided both critical feedback on the current TOAS software applied analysis techniques and valuable insights on the vary enabling technologies being applied. The bibliometric analyses methodologies followed by the authors have surfaced unspoken or undocumented problems, which were encountered in previous analyses and methodically remedied. The authors each conducted independent evaluations of the NLP and CL abstract files; with minor exceptions, the inferences documented in this paper represent common conclusions. Peer review of our approaches has identified best practices and personal preferred processes' oversights or weaknesses, all of which will be considered as we continue to develop the TOAS system.

Search Technology, Inc., in collaboration with Intelligent Information Services Corporation (IISC) and the Georgia Institute of Technology, appear fully committed to commercializing the current software. Beta users have been enlisted to promote product interest and identify market-desired features/functions. We hope to create a system that interfaces with the user through a Windows-based menu-driven environment. Once tailored by the user to the database field formats of interest, the TOAS could semi-automatically identify contributing technologies within the topic area. The user could then click on any identified sub-grouping of terms to create transitional files and obtain

record field summary listings and comparative analyses to gain valuable insights on social, intellectual and temporal indicators.

We hope to expand the Technology Opportunities Analysis System (TOAS) functionality, with DARPA sponsorship, to include information synthesis. The TOAS capabilities could be greatly enhanced by the addition of an intelligent database recognition front end, quantitative graphical output displays, and the incorporation of data mining tools to complement the current text analysis capabilities. The intelligent front end could enable rapid tailoring of the TOAS to any fixed field textual database. Graphical output will ease user recognition of revealed patterns within the documentation analyzed. The integration of data mining tools would allow pattern recognition and identification in quantitative databases, such as the Army Operation and Support Management Information System (OSMIS). In the future, the system could be programmed to link identified data mined technology requirements (i.e., quantitative data clusters) to research literature summaries (i.e., textual factor groupings) on potential solutions. The evolution of virtual organizations could be supported by this capability. Obviously, both tactical and strategic goals could be supported. Whether this vision becomes reality will depend on adequate time and resource commitments.

References:

1. Cunningham, S. (1996), The Content Evaluation of British Scientific Research, Science Policy Research Unit: Brighton, U.K.
2. Deerwester, S. and S.T. Dumais, G.W. Furnas, T.K. Landauer, D. Harshman (1990), "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, Vol. 41 No. 6, p. 391-407.
3. Kostoff, R. N. (1993), "Database Tomography for Technical Intelligence," Competitive Intelligence Review Vol. 5 No. 1, p. 1-9.
4. Kostoff, R. N. (1994), "Database Tomography: Origins and Applications," Competitive Intelligence Review, Vol. 5 No. 1, p.48-55.
5. Porter, Alan L. and Detampel, Michael, J. (1995), "Technology Opportunities Analysis," Technology Forecasting and Social Change, Vol. 49, p. 237-255.
6. Press, W. H. and B. P. Flannery, S.A. Teukolsky, W. T. Vetterling (1986), Numerical Recipes in C, Cambridge University Press: Cambridge.
7. Watts, Robert J. and Porter, Alan L. (to appear), "Innovation Forecasting," Technology Forecasting and Social Change.