Science and technology text mining::
Text Mining of the Journal *Cortex*

Ronald N. Kostoff[1], Henry A. Buchtel (2,3), John Andrews (4), Kirstin M. Pfeil[5]

1, 4, 5Office of Naval Research, 2 VA Ann Arbor Healthcare System, 3 University of Michigan Departments of Psychiatry and Psychology.

Correspondence to:

Ronald N. Kostoff, Ph.D.
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

ABSTRACT

Background: The stated mission of *Cortex* is "the study of the inter-relations of the nervous system and behavior, particularly as these are reflected in the effects of brain lesions on cognitive functions." The purpose of this report is to explore the relationship between the stated mission and the executed mission as reflected by the characteristics of papers published in *Cortex*. In addition, we examine whether the results and conclusions of an analysis of this kind are affected by the level of description of the published papers.

Objectives:
A) Identify characteristics of contributors to *Cortex*;
B) Identify characteristics of those who cite *Cortex*;
C) Identify recurring themes;
D) Identify the relationships among the recurring themes;
E) Compare recurring themes and determine their relationships to the mission of *Cortex*;
F) Identify the sensitivity of these results to the level of description of the *Cortex* papers used as the source database.
G) Compare *Cortex* characteristics with those of *Neuropsychologia*, another Europe-based international neuropsychology journal.

Methods: Text mining (extraction of useful information from text) was used to generate the characteristics of the journal *Cortex*. Bibliometrics provided the *Cortex* contributor infrastructure (author/ organization/ country/ citation distributions), and computational linguistics identified the recurring technical themes and their inter-relationships. Citation mining (the integration of citation bibliometrics and text mining) was used to profile the research user community. Four levels of published article description were compared for the analysis: Full Text, Abstract, Title, Keywords.

Results: and Conclusions: Highly cited documents were compared among *Cortex, Neuro-psychologia,* and *Brain*, and a number of interesting parametric trends were observed. The characteristics of the papers that cite *Cortex* papers were examined, and some interesting insights were generated. Finally, the document clustering taxonomy showed that papers in *Cortex* can be reasonably divided into four categories (papers in each category in parenthesis): Semantic Memory (151); Handedness (145); Amnesia (119); and Neglect (66).

It is concluded that *Cortex* needs to take steps to attract a more diverse group of contributors outsied its continental Western European base if it wishes to capture a greater share of seminal neuropsychology papers. Further investigation of the critical citation differences reported in the report is recommended.

INTRODUCTION

The stated mission of *Cortex* is "the study of the inter-relations of the nervous system and behavior, particularly as these are reflected in the effects of brain lesions on cognitive functions." (See Journal Title Page) The aim of this report is to examine the relationship between the stated mission and the executed mission as reflected by the characteristics of papers published in the journal. This was done by determining the technical and thematic characteristics of papers published in Cortex, and their inter-relationships as expressed by the categories in different taxonomies. In addition, we set out to ascertain the infrastructure (authors, institutions, countries) underlying the papers published in *Cortex*, as well as the infrastructure and technical focus of the community of authors who cite papers published in *Cortex*. Finally, we were interested in determining whether the results and conclusions about the technical themes and their relationships differ according to the level of information contained in the specific record field analyzed (Keywords, Titles, Abstracts, or Full Text).

For the past decade, the first author has been developing ways to obtain the above types of information from large bodies of unstructured or semi-structured text (1-8). These processes are collectively known as text mining (9-14). They consist of three generic components: information retrieval, information processing, and information integration. It was decided to apply text mining to obtain the different perspectives on Cortex outlined above. An iterative query development process is usually used for this kind of task (15), but this is not needed when analyzing a database of papers published in a particular journal. The main focus of the study was the application of information processing and information integration to a subset of the papers published in Cortex.

METHODS

There are four components of the specific approach selected: database selection, bibliometrics analysis, computational linguistics analysis, and citation mining. Each will be outlined.

I. Database Selection

Two databases were used for the study. The first database was the Web version of the Science Citation Index (SCI) (16), which consisted of all *Cortex* records from 1991-mid-2001 classified in the SCI as articles. Four hundred ninety-four records were retrieved, of which 481 were full articles with abstracts. Most of the records included authors, titles, author addresses, author keywords, abstract narratives, and references cited.

The second database consisted of all of the 203 full text *Cortex* articles published from 1997 to 2000. These articles were supplied by the publisher in electronic format.

II. Bibliometric Analysis (4-8)

The purpose of the bibliometrics analysis is to quantify the basic technical infrastructure of *Cortex*. This quantification is obtained through counting items such as authors, institutions, countries, and citations. While the quantification procedure is straight-forward, its interpretation can be quite complex.

The bibliometrics section has two components: Publication Bibliometrics (e.g., prolific authors and numbers of papers published); and Citation Bibliometrics. These are compared with similar results from the journal Neuropsychologia, and in one case, results from the journal Brain are included as well.

III. Citation Mining (19)

Citation mining integrates citation bibliometrics and computational linguistics. Its purposes are to profile the documented user community, and to show the technical disciplines into which the cited research areas are evolving. In citation mining, a sample of papers describing the research area is selected, and all papers in the SCI that cite the sample papers are retrieved. Bibliometrics and computational linguistics are performed on this sample. The bibliometrics displays characteristics of the citing community, and the computational linguistics portrays the technical thrusts (and interrelationships) of the citing disciplines.

The sample selected consists of all articles published in Cortex in 1993-1994. There were a total of 73 papers selected. Over 1300 separate citing articles were retrieved. These 1300 citing articles were citation mined.

IV. Computational Linguistics Analysis (1-8)

The purpose of the computational linguistics is to use the quantification of text patterns to identify the technical themes of the database, the relationship among those themes, and the relationship between the themes and the technical infrastructure revealed by this bibliometric analysis. The approach used in the present study was to count phrase combinations that co-occurred within bounded domains (e.g., Abstracts, specified numerical windows), and group documents that appeared in thematic clusters.

B. Taxonomy Generation: Statistical Clustering (11, 22)

General analytic approach

For the long-term *Cortex* analysis, the taxonomy of the Abstract field database covering *Cortex* papers from 1991-2001 was generated. Past text mining studies have used a variety of approaches to identify the main technical themes in the database. These include extracting key phrases and manually assigning them to categories; extracting key phrases and assigning them with statistical computer algorithm, using factor analyses and multi-link clustering; and grouping documents based on text similarity.

While factor analysis, multi-link phrase clustering, and document clustering were used for the present study, only document clustering will be reported in the main text,. The other computational linguistics approaches and results are presented in the Appendices. The three techniques provided complementary perspectives on the structure of the Cortex literature. For the total SCI database, document clustering was performed using the Abstracts text only. In document clustering, documents are combined into groups based on their text similarity. Document clustering yields number of documents in each cluster directly, a proxy metric for level of emphasis in each taxonomy category.

Different document clustering approaches exist [39-48]. The approach presented in this section is based on a partitional clustering algorithm [49-50] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements.

CLUTO requires specification of the number of clusters desired. Cluster runs (of the total SCI database) of 32 clusters were generated. CLUTO also agglommorated the 32 clusters into a hierarchical tree (taxonomy) structure, and this taxonomy is presented in the clustering sections.

RESULTS

I. Publication Bibliometrics

The first group of metrics consists of counts of papers published by different entities e.g., authors, countries in which the work was carried out). These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred due to these papers' publication in the (typically) high caliber of journals accessed by the SCI.

A. Prolific authors

Table 1 lists the twenty most prolific first authors from Cortex and Neuropsychologia in this sample.

TABLE 1

CORTEX		NEUROPSYCHOLOGIA	
AUTHOR	**FREQ**	**AUTHOR**	**FREQ**
MAYES—AR (UK)	11	HODGES—JR (UK)	19
CARLESIMO, Giovanni A (ITALY)	10	COWEY, Alan (UK)	16
HEILMAN, Kenneth M (USA)	10	GRAFMAN, Jordan (USA)	16
PILLON, Bernard (FRANCE)	10	HEILMAN, Kenneth M (USA)	15
CALTAGIRONE, Carlo (ITALY)	9	MILNER, Brenda (CANADA)	15
DUBOIS, B (FRANCE)	9	RUGG, Michael D (UK)	14
AGID, Yves (FRANCE)	8	WARRINGTON, Elizabeth K (UK)	13
DENES, G (ITALY)	8	ROBBINS, Trevor W (UK)	12
DERENZI, Ennio (ITALY)	8	ROBERTSON, IH (UK)	12
GRAFMAN, Jordan (USA)	8	BRADSHAW, John L (AUSTRALIA)	11
WARRINGTON, Elizabeth K (UK)	8	CORBALLIS, Michael C (NZ)	11
CAPITANI, Erminio (ITALY)	7	FRITH, CD (UK)	11
SIRIGU, Angela (FRANCE)	7	GAZZANIGA, Michael S (USA)	11
ANNETT, M (UK)	6	PARKIN, Alan J (UK)	11
BASSO, Anna (ITALY)	6	BRYDEN, M Philip (CANADA)	10
CIPOLOTTI, L (UK)	6	DRIVER, Jon (UK)	10
PIZZAMIGLIO, Luigi (ITALY)	6	MAYES, Andrew R (UK)	10
SABBADINI, Maurizio (ITALY)	6	PATTERSON, K (UK)	10
UMILTA, Carlo (ITALY)	6	DOLAN, Ray J (UK)	9
ADAIR, JC (USA)	5	FARAH, Martha J (USA)	9

Of the twenty most prolific authors in Cortex, nine are from Italy, four are from the UK, four are from France, and three are from the USA. For Neuropsychologia, twelve are from the UK, four are from the USA, two are from Canada, one is from Australia, and one is from New Zealand. There are four names in common between the two lists (Mayes, Warrington, Heilman, Grafman). The first two authors are from the UK; the latter two are from the USA. The country distributions of the top twenty most prolific

authors are very different, and different from those of other recent text mining studies performed by the first author. Almost half of the top performers in *Cortex* are from Italy, and Western Europe generally. The *Neuropsychologia* top performers are centered in the UK primarily, and in the countries of the TTCP (The Technical Cooperation Program) totally.

B. Prolific organizations

Table 2 lists the twenty most prolific institutions. It should be noted that many different organizational components may be included under a single organizational heading (e.g., The University of Milan could include the Neurology Department, Neuropsychology Department, Neuroscience Department, etc.).

TABLE 2

CORTEX		NEUROPSYCHOLOGIA	
INSTITUTION	**FREQ**	**INSTITUTION**	**FREQ**
UNIV MILAN	18	UNIV OXFORD	41
UNIV PADUA	17	UNIV CAMBRIDGE	34
HOP LA PITIE SALPETRIERE	13	UNIV COLL LONDON	34
UNIV FLORIDA	11	MRC	32
IRCCS S LUCIA	10	MCGILL UNIV	30
UNIV MODENA	10	HARVARD UNIV	22
BOSTON UNIV	9	UNIV FLORIDA	20
HOP HENRI MONDOR	9	UNIV CALIF DAVIS	20
INSERM	9	UNIV CALIF LOS ANGELES	20
UNIV COLL LONDON	8	UNIV MILAN	20
CTR PAUL BROCA	8	UNIV CALIF SAN DIEGO	19
VET ADM MED CTR	8	NATL HOSP NEUROL & NEUROSURG	18
INST PSYCHIAT	8	HOP LA PITIE SALPETRIERE	17
UNIV CALIF LOS ANGELES	7	UNIV PENN	17
UNIV ABERDEEN	7	UNIV MONTREAL	16
UNIV PENN	7	UNIV TORONTO	16
UNIV QUEENSLAND	7	UNIV PADUA	16
IRCCS	7	BOSTON UNIV	16
NATL HOSP NEUROL & NEUROSURG	7	UNIV DUSSELDORF	15
UNIV ST ANDREWS	6	IRCCS	15

A number of observations from Table 2 follow. First, a slight majority of the twenty most prolific Cortex institutions are in academic settings (55%); the others are hospitals (25%) and research institutions (20%). It should be pointed out, however, that clinicians and researchers working in hospitals such as Hôpital La Pitie Salpêtriére (Paris) and The National Hospital for Neurology and Neurosurgery (Queen Square, London UK) usually have academic affiliations (London University, in the case of the National Hospital), so the institutional name may not identify the clinical versus academic status of the authors completely.

Second, a substantial majority of the twenty most prolific Neuropsychologia institutions are in academic settings (85%); the others are hospitals (10%) and research institutions

(5%). Third, of the five common institutions located in the upper part of the *Cortex* column, three are located in the lower part of the *Neuropsychologia* column.

Finally, seven of the top ten in the Cortex column are from continental Western Europe, with the dominant two institutions being from Italy. Contrast this with the Neuropsychologia column, where nine of the top ten are from the predominantly English speaking countries of USA, UK, and Canada.

C. Prolific countries

Table 3 lists the twenty most prolific countries. While the USA has reasonably similar representation in both journals, Italy is represented about 2.5 times as much in Cortex, whereas England is represented about 50% more in Neuropsychologia. The prolific country results, especially at the top of the table, track the prolific author results. However, as will be shown in the later analysis of most and least cited papers published in these journals, the prolific author/ country results do not track the most cited paper results as well. In particular, the most cited papers published in Cortex, in the 1998-1999 sample examined, come from continental Western Europe, mainly Italy and France, and the USA and UK are not represented. The most cited papers published in Neuropsychologia come from the English-speaking countries, mainly UK and USA, reflecting the most prolific authors/ countries.

CORTEX			**NEUROPSYCHOLOGIA**		
COUNTRY	**FREQ**	**FRAC**	**COUNTRY**	**FREQ**	**FRAC**
USA	241	0.2726	USA	784	0.3286
ITALY	179	0.2025	ENGLAND	491	0.2058
ENGLAND	122	0.138	ITALY	200	0.0838
FRANCE	108	0.1222	CANADA	193	0.0809
GERMANY	42	0.0475	FRANCE	156	0.0654
JAPAN	40	0.0452	GERMANY	155	0.065
CANADA	29	0.0328	AUSTRALIA	82	0.0344
SCOTLAND	23	0.026	SCOTLAND	53	0.0222
AUSTRALIA	23	0.026	NETHERLANDS	40	0.0168
BELGIUM	19	0.0215	SWITZERLAND	36	0.0151
NETHERLANDS	12	0.0136	JAPAN	35	0.0147
FINLAND	11	0.0124	BELGIUM	28	0.0117
AUSTRIA	7	0.0079	SPAIN	26	0.0109
SWEDEN	7	0.0079	ISRAEL	24	0.0101
ISRAEL	4	0.0045	NORWAY	17	0.0071
SPAIN	4	0.0045	SWEDEN	16	0.0067
GREECE	4	0.0045	FINLAND	15	0.0063
THAILAND	3	0.0034	NEW ZEALAND	14	0.0059
BRAZIL	3	0.0034	WALES	14	0.0059
SWITZERLAND	3	0.0034	DENMARK	7	0.0029

TABLE 3

In both journals, the dominance of a handful of countries is clearly evident. In Cortex, two countries, USA and Italy, are represented in 48% of the author address listings, while in Neuropsychologia, two countries, USA and UK, are represented in 53% of the author address listings.

II. Citation Bibliometrics

The second group of metrics presented is counts of citations to papers published by different entities. While citations are often used as impact or quality metrics (26), much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers (25-27).

The references (citations) in all the 494 retrieved papers were aggregated. Each paper citation was divided into author, year, and journal fields, and those cited most frequently were identified. The data were accumulated and presented in order of decreasing frequency. A small percentage of any of these categories received large numbers of citations.

A. Most cited first authors

Table 4 lists the twenty most cited first authors.

TABLE 4

CORTEX		NEUROPSYCHOLOGIA	
AUTHOR	**FREQ**	**AUTHOR**	**FREQ**
DERENZI E	286	WARRINGTON EK	411
WARRINGTON EK	250	MILNER B	396
ANNETT M	158	POSNER MI	386
SHALLICE T	148	SHALLICE T	284
BENTON AL	107	FARAH MJ	256
MILNER B	100	DERENZI E	250
BISIACH E	94	HEILMAN KM	241
WECHSLER D	90	SQUIRE LR	229
TULVING E	86	WECHSLER D	225
SQUIRE LR	82	BISIACH E	218
HEILMAN KM	76	KINSBOURNE M	216
KINSBOURNE M	76	SCHACTER DL	182
FARAH MJ	74	PETRIDES M	181
SCHACTER DL	74	TULVING E	176
GESCHWIND N	73	SERGENT J	175
KOPELMAN MD	73	KIMURA D	174
KAPUR N	69	KOSSLYN SM	171
BRYDEN MP	69	BENTON AL	170
NELSON HE	67	BRADSHAW JL	170
CARAMAZZA A	65	DAMASIO AR	169

There are thirteen authors in common between the two columns. Apart from Annett (*Cortex*) and Posner (*Neuropsychologia*), the differences are seen in the lower portion of the columns. These two authors reflect the areas of emphasis of the two journals, with handedness being the most most frequent category of articles in *Cortex* and cognition and attention being a frequent subject in *Neuropsychologia*. There is modest overlap between the most prolific authors and most cited first authors in both journals, being 20% in Cortex and 25% in *Neuropsychologia*. This modest overlap has been found in most text mining studies performed by the first author. It may be due to the time lag between an author's seminal works and present activity, the difficulty in being prolific and producing seminal papers, or prolific authors not being listed as first authors in papers that are co-authored by students and junior colleagues.

The latter cause may be significant. For example, the ten most recently published papers (all journals) of the five most prolific authors in *Neuropsychologia* (from Table 1), and the five most prolific authors in Cortex (excluding Heilman, who is captured in the *Neuropsychologia* group) were examined. These ten authors in Table 1 were first authors in only 9 of the 100 total papers, similar to the experience for most of the other text mining studies conducted by the first author. For those prolific authors who are rarely first authors, the chances that they will accumulate significant first author citations are low.

B. Most cited documents

The twenty most cited documents in *Cortex* and in *Neuropsychologia* are listed in Table 5. There are eleven documents in common between the two lists. The top ten document citations center around 1980. For *Cortex*, the most cited papers and books range from 1966 to 1991 (from 25-0 years before the start of the survey), with the median being 1980. For *Neuropsychologia*, the most cited papers and books range from 1964-1990, with the median being 1982.

Because of space limitations, only the ten most cited *Cortex* papers will be summarized briefly. The Folstein paper describes the popular MMSE (Mini-Mental State Exam). The Oldfield paper describes a paper-and-pencil questionnaire for measuring the degree of right- and left-handedness. The Warrington reference refers to the Warrington Recognition Memory Test for Faces, which uses fifty target faces, but includes a considerable amount of non-facial information in its stimuli. The Snodgrass paper provides stimuli for use in studies in which the complexity and other characteristics of the stimulus need to be known. The Shallice book describes the connection between cognitive psychology and neuropsychology. The McKhann paper reports the consensus statement defining the characteristics of possible and probable Alzheimer's Disease The Shallice and Evans paper describes the tendency of patients with frontal lobe lesions to make errors when asked to make estimates of sizes and frequencies that are not generally known with certainty (e.g., how many camels are there in Holland). The Nelson paper introduced a simplified version of the Wisconsin Card Sorting Test appropriate for elderly patients or patients likely to be fatigued by the original version.. The Albert paper

concerns the dissociation of neglect as a disorder from visual field defects. The Derenzi paper addresses normative data and the screening power of a shortened version of the Token Test. The Annett book addresses the demographics and putative genetics of handedness. Overall, the most cited Cortex reference documents tend to focus on clinical behavioral tests, as opposed to surgical experiments, invasive diagnostic experiments, animal laboratory tests, or even non-invasive experiments, to any significant degree.

Taken together, this list of most cited papers has no single theoretical or conceptual theme, but papers on classification, methods and tests are clearly represented more often than one would expect by chance. In particular, the paper by Folstein et al. (1975) is frequently cited because the test described in that paper (MMSE) is used across disciplines as an approximate measure of cognitive status. The test requires little training to administer and covers several domains relevant to evidence of cognitive decline (orientation to person and place, memory, comprehension, calculations and internal mental manipulations, and the like). The exceptions to the "methods" theme are the books by Shallice and by Annett and the paper by Shallice and Evans (1978). While a task is described in the Shallice and Evans paper, that particular version of the test is not in common use, but the concept of executive function described in this paper continues to be important in neuropsychology (a modified version of the method, with norms, was published later by Axelrod et al., 1994 (28)).

TABLE 5

CORTEX

DOCUMENT	FREQ
FOLSTEIN MF, 1975, J PSYCHIAT RES, V12, P189	52
OLDFIELD RC, 1971, NEUROPSYCHOLOGIA, V9, P97	41
WARRINGTON EK, 1984, RECOGNITION MEMORY T	35
SNODGRASS JG, 1980, J EXPT PSYCHOL HUMAN, V6, P174	34
SHALLICE T, 1988, NEUROPSYCHOLOGY MENT	34
MCKHANN G, 1984, NEUROLOGY, V34, P939	33
SHALLICE T, 1978, CORTEX, V14, P294	31
NELSON HE, 1976, CORTEX, V12, P313	30
ALBERT ML, 1973, NEUROLOGY, V23, P658	24
DERENZI E, 1978, CORTEX, V14, P41	23
ANNETT M, 1985, LEFT RIGHT HAND BRAI	23
WECHSLER D, 1981, WECHSLER ADULT INTEL	22
DERENZI E, 1987, CORTEX, V23, P575	21
WARRINGTON EK, 1984, BRAIN, V107, P829	18
ANNETT M, 1970, BRIT J PSYCHOL, V61, P303	18
WARRINGTON EK, 1991, VISUAL OBJECT SPACE	17
BENTON AL, 1983, CONTRIBUTIONS NEUROP	17
MILNER B, 1971, BRIT MEDICAL B, V27, P272	17
BRUCE V, 1986, BRIT J PSYCHOL, V77, P305	16
LURIA AR, 1966, HIGHER CORTICAL FUNC	16

NEUROPSYCHOLOGIA

DOCUMENT	FREQ
OLDFIELD RC, 1971, NEUROPSYCHOLOGIA, V9, P97	141
FOLSTEIN MF, 1975, J PSYCHIAT RES, V12, P189	98
SHALLICE T, 1988, NEUROPSYCHOLOGY MENT	78
SNODGRASS JG, 1980, J EXPT PSYCHOL HUMAN, V6, P174	75
WECHSLER D, 1981, WECHSLER ADULT INTEL	65
POSNER MI, 1984, J NEUROSCI, V4, P1863	58
MCKHANN G, 1984, NEUROLOGY, V34, P939	55
TALAIRACH J, 1988, COPLANAR STEREOTAXIC	55
NELSON HE, 1976, CORTEX, V12, P313	54
WARRINGTON EK, 1984, RECOGNITION MEMORY T	50
POSNER MI, 1990, ANNU REV NEUROSCI, V13, P25	49
MESULAM MM, 1981, ANN NEUROL, V10, P309	46
POSNER MI, 1980, Q J EXPT PSYCHOL, V32, P3	44
WARRINGTON EK, 1984, BRAIN, V107, P829	44
MILNER B, 1971, BRIT MEDICAL B, V27, P272	43
KUCERA H, 1967, COMPUTATIONAL ANAL P	40
MILNER B, 1964, FRONTAL GRANULAR COR, P313	39
ALBERT ML, 1973, NEUROLOGY, V23, P658	39
STUSS DT, 1986, FRONTAL LOBES	38
GOLDMANRAKIC PS, 1987, HDB PHYSL 1, V5, P373	37

C. Most cited journals

Table 6 lists the top twenty most cited journals and their frequencies.

TABLE 6

CORTEX		NEUROPSYCHOLOGIA	
JOURNAL	FREQ	JOURNAL	FREQ
NEUROPSYCHOLOGIA	1592	NEUROPSYCHOLOGIA	5293
CORTEX	1342	BRAIN	2228
BRAIN	807	CORTEX	1690
BRAIN LANG	572	NEUROLOGY	1204
NEUROLOGY	515	BRAIN LANG	1062
J NEUROL NEUROSUR PS	447	SCIENCE	916
COGNITIVE NEUROPSYCH	406	NATURE	898
BRAIN COGNITION	403	J NEUROSCI	869
ARCH NEUROL-CHICAGO	353	J NEUROL NEUROSUR PS	855
J CLIN EXP NEUROPSYC	270	BRAIN COGNITION	832
SCIENCE	221	ARCH NEUROL-CHICAGO	777
NATURE	185	COGNITIVE NEUROPSYCH	728
ANN NEUROL	160	EXP BRAIN RES	684
J EXP PSYCHOL LEARN	158	J NEUROPHYSIOL	636
BRIT J PSYCHOL	155	J COGNITIVE NEUROSCI	605
J NEUROSCI	138	J CLIN EXP NEUROPSYC	535
PSYCHOL REV	117	J EXP PSYCHOL LEARN	522
PSYCHOL BULL	115	PSYCHOL REV	478
COGNITION	114	J EXP PSYCHOL HUMAN	469
PERCEPT MOTOR SKILL	109	ANN NEUROL	459

The journals at the top of the lists are mainly neurology and neuropsychology journals, with the exception of the general science journals *Science* and *Nature*. The bottom of the lists includes psychology journals that publish papers relevant to neuropsychology. In agreement with the emphasis on clinical findings and their significance, there are no highly cited journals specializing in basic genetics, biology or biochemistry. The first entries on the most cited journals lists that reflect other fields of science are, for *Cortex*, *Journal of the Acoustics Society of America (JASA)* and *American Journal of Medical Genetics* (numbers 110 and 176 on the list of journals, respectively), and, for *Neuropsychologia*, *JASA*, *Biological Cybernetics*, and *Cell* (numbers 71, 230, and 459 on the list of journals, respectively).

There are sixteen journals in common between the two lists. Add to this the majority commonality in Cortex and Neuropsychologia of most cited authors and most cited documents shown previously, and it can be concluded that both journals draw heavily upon the same intellectual heritage. Given the many similarities in intellectual heritage, and the modest similarities in production demographics (prolific authors, institutions), how do the papers published in the two journals impact the larger technical community? This question is partially answered in the next section.

D. Citation Comparison among *Cortex*, *Neuropsychologia*, and *Brain*

To further compare citations among papers published in the three top cited journals above, *Cortex*, *Neuropsychologia*, and *Brain*, the following experiment was run. All articles published in Cortex, Neuropsychologia, and Brain in the years 1998-1999 were retrieved from SCI. There were 110 *Cortex* articles, 278 *Neuropsychologia* articles, and 341 *Brain* articles. Then, the ten most cited articles from each retrieval (the citations from each paper used for the tabulation of most and least cited are those listed in the SCI Times Cited field, and are the total citations received by each paper from all other papers in the SCI) were extracted, as well as the ten least cited articles, and various characteristics compared. The results are shown in Table 7

TABLE 7

	CORTEX		**NEUROPSY**		**BRAIN**	
	MOST CITED	LEAST CITED	MOST CITED	LEAST CITED	MOST CITED	LEAST CITED
# AUTH						
AVER	3.9	2.8	5.2	2.6	7.1	4.6
MEDIAN	4	3	5	1	7.5	4.5
# REFS						
AVER	46.3	28	52.5	26.8	68.3	42.4
MEDIAN	49	29.5	49	26	62.5	35
# CITES						
AVER	21	0.8	71.3	0	166.8	2.8
MEDIAN	18.5	1	67.5	0	157	3
ORG	5	4	2	4	8	2
INST	5	4	2	4	8	2
UNIV	5	6	8	6	2	8
COUNTRY	4 ITALY	2 ITALY	4 UK	5 USA	5 UK	3 JAPAN
	3 FRANCE	2 USA	4 USA	2 ITALY	2 USA	1 USA
	1 AUSTRIA	2 GERMANY	1 ITALY	1 NZ	2 CANADA	1 UK
	1 BELGIUM	2 JAPAN	1 CANADA	1 NETH	1 GERMANY	1 FRANCE
	1 GERMANY	1 NETH		1 AUSTRALIA		1 ITALY
		1 AUSTRALIA				1 CANADA
						1 GERMANY
						1 NETH
TYPE						
BEHAV	8		4			
SURGERY			1		2	
DIAG-NI	2		5		7	
DIAG-INV					1	

CODE:TYPE
BEHAV=CLINICAL BEHAVIOR STUDIES
SURGERY=SURGICAL INTERVENTIONS
DIAG-NI=NON-INVASIVE DIAGNOSTIC TESTS
DIAG-INV=INVASIVE DIAGNOSTIC TESTS

The most cited articles in Neuropsychologia are cited, on average, more than three times as often as the most cited articles in Cortex, and the most cited articles in Brain are cited, on average, more than twice as often as the most cited articles in Neuropsychologia.

Second, the most cited papers have more authors than the least cited, in all three journals, and the effect is most pronounced in *Neuropsychologia*. Additionally, the average number of authors increases with the average number of citations, ranging from about four authors of the most cited *Cortex* papers to about seven authors of the most cited *Brain* papers.

Third, the most cited papers have substantially more references than the least cited, in both journals, and the effect is most pronounced in *Neuropsychologia*. Additionally, the average number of citations increases with the average number of references (an effect observed by the first author in recent unpublished text mining studies), ranging from about 46 references in the most cited *Cortex* papers to about 68 references in the most cited Brain papers.

Fourth, there is no clear overall trend in citations as a function of institutional representation. The institution/ (institution + university) ratio (where institution in the table cells should be interpreted as any non-university organization; e.g., research laboratory, clinic, hospital, company) for most cited papers starts at 0.5 for Cortex, drops to 0.2 for *Neuropsychologia*, and increases sharply to 0.8 for *Brain*. This ratio for least cited papers starts at 0.4 for both *Cortex* and *Neuropsychologia*, and decreases to 0.2 for *Brain*. Its most dramatic change is from 0.8 for the most cited *Brain* papers to 0.2 for the least cited *Brain* papers.

Fifth, the most cited papers in *Cortex* are all from continental Western Europe, with heavy representation from Italy and France, while the least cited papers in *Cortex* represent four different continents. The most cited papers in *Neuropsychologia* are, with the exception of Italy, from the UK and North America (with heavy representation from the UK and USA), while the least cited papers have more representation from Western Europe but none from the UK. The most cited papers in *Brain* are from the major English-speaking countries, whereas the least cited are scattered around Western Europe, Asia, and North America.

Sixth, there is a distinct shift in type of study (the bottom of Table 7) in proceeding from Cortex to Neuropsychologia to Brain. Clinical behavioral studies, many of them essentially case studies, predominate the most cited Cortex papers. There are only two papers characterized as Diagnostic-Non-Invasive (e.g., PET, MRI, etc). Neuropsychologia has more of a balance between Behavioral and Diagnostic-Non-Invasive in its ten most cited papers. Brain shows a heavy emphasis on Diagnostic-Non-Invasive (7/10), two papers on surgical procedures, and one on Diagnostic-Invasive. Based on reading Abstracts from each of these journals, the types as represented in the top ten most cited articles roughly approximate the types of papers published overall.

Thus, as citations increase in absolute amounts, the study type transitions from the clinically oriented behavioral focus to the correlates with more objective measurements.

III. Citation Mining

A. Bibliometrics

For the 73 *Cortex* sample papers published in 1993-94, there were over 1300 citing papers. There were a total of 1238 citing authors, the top ten of which are shown in Table 22. Interestingly, there is no overlap between the list of most often cited authors and authors who most frequently cite papers published in Cortex.

TABLE 22 – AUTHORS OF CORTEX CITING PAPERS – TOP 10

RANK	#RECORDS	AUTHOR
1	13	Markowitsch, Hans J. (Germany)
2	12	Grafman, Jordan (USA)
3	12	Sirigu, Angela (France)
4	10	Pillon, Bernard (France)
5	9	Tulving, Endel (Canada)
6	9	Doty, Robert (USA)
7	9	Kessler, Josef (Germany)
8	8	Daum, Irene (Germany)
9	8	Dubois, Bruno (France)
10	8	Agid, Yves (France)

As shown in Table 23, the 73 *Cortex* sample papers were cited in 152 different journals. Most of the citing papers were published in *Cortex*. It is clear that papers published in *Cortex* are of fundamental importance to subsequent papers to be published there. To determine whether this has its origin in a "niche" occupied by the journal or instead implies a degree of narrowness would require a different kind of analysis, one requiring a close analysis of the contents of the papers citing the *Cortex* articles.

TABLE 23 – JOURNALS PUBLISHING PAPERS THAT CITE CORTEX

Affiliation (Journal)
Total Journals: 152
Top 10

	# Records	Journal
1	521	*Cortex*
2	24	*Neuropsychologia*
3	17	*Cognitive Neuropsychology*
4	16	*Brain*
5	14	*Journal of Cognitive Neuroscience*
6	12	*Neurocase*
7	11	*Neuropsychology*
8	11	*Brain and Cognition*
9	11	*Brain and Language*
10	10	*Neuroimage*

As shown in Table 24, the citing authors came from 460 institutions. Most of the highest frequency institutions are Western European.

TABLE 24 – ORGANIZATIONS OF CORTEX CITING AUTHORS

Affiliation (Organization)
Total Organizations: 460
Top 10

	# Records	*# Instances*	*Affiliation (Organization)*
1	16	24	Hôpital La Pitie Salpêtriére (France)
2	13	17	University of Bielefeld (Germany)
3	12	18	Medical Research Council (UK)
4	12	24	University of Tübingen (Germany)
5	11	19	University College, London (UK)
6	11	24	University of Pennsylvania (USA)
7	10	15	University of Toronto (Canada)
8	8	13	INSERM (France)
9	8	10	Institute for Cognitive Science (France)
10	8	11	University of Cambridge (UK)

As shown in Table 25, the highest frequency citing countries are the USA and Western Europe. Given the dearth of US institutions in the top ten (Table 24), this means that the participation of US institutions is relatively widespread.

TABLE 25 – COUNTRIES OF CORTEX CITING AUTHORS

Affiliation (Country)
Total Countries: 30
Top 10

	# Records	*# Instances*	*Affiliation (Country)*
1	142	406	USA
2	91	225	UK
3	57	154	Germany
4	55	148	France
5	33	73	Italy
6	30	66	Canada
7	16	32	Belgium
8	14	40	Japan
9	1	20	Australia
10	1	14	Netherlands

As shown in Table 26, 8531 different authors were cited in the papers that cited the Cortex sample papers.

TABLE 26 – FIRST AUTHORS CITED BY CORTEX CITING AUTHORS

Cited First Authors
Total Cited First Authors: 8531
Top 10

	# Records	*# Instances*	*Cited First Author*
1	131	226	Shallice, Tim (UK)
2	108	175	Kapur, Nirander (UK)
3	98	245	Tulving, Endel (Canada)
4	93	189	DeRenzi, Ennio (Italy)
5	91	187	Squire, Larry (USA)
6	79	144	Warrington, Elizabeth (UK)
7	78	213	Schacter, Daniel (USA)
8	74	114	Wechsler, David
9	69	106	Damasio, Antonio
10	64	108	Haxby, James V

As shown in Table 27, the *Cortex* sample citing papers cited a total of 17550 different books and papers. The only *Cortex* paper in the top ten was the paper by Nirandar Kapur. The top ten citations are mainly to journal articles (8/10) and mainly to work since 1990. The exceptions are the Folstein et al. (1975) paper, already discussed (MMSE), the Shallice book (1988) and an atlas (Talairach & Tournoux (1988), which is used for localizing brain areas in imaging studies.

TABLE 27 – PAPERS CITED BY CORTEX CITING AUTHORS

Cited Papers
Total Cited Books and Papers: 17550
Top 10

	# Records	# Instances	*Cited Paper*
1	78	78	Kapur N, 1995, *CORTEX*, V31, P99
2	45	45	Sergent J, 1992, *BRAIN*, V115, P15
3	41	41	Shallice T, 1988, *NEUROPSYCHOLOGY MENT*
4	40	40	Haxby JV, 1996, *P NATL ACAD SCI USA*, V93, P922
5	39	39	Shallice T, 1994, *NATURE*, V368, P633
6	38	38	Tulving E, 1994, *P NATL ACAD SCI USA*, V91, P2016
7	37	37	Shallice T, 1991, *BRAIN*, V114, P727
8	35	35	Talairach J, 1988, *COPLANAR STEREOTAXIC*
9	33	33	Grady CL, 1995, *SCIENCE*, V269, P218
10	32	32	Folstein MF, 1975, *J PSYCHIAT RES*, V12, P189

Finally, there were 3769 journals cited by the citing papers of the sample *Cortex* papers. The top ten are shown in Table 29. Based on all the bibliometrics results, citation mining and non-citation mining, there appears to be a symbiotic relationship among *Cortex*, *Neuropsychologia*, and *Brain*. Of the journals in the list, these three are the most relevant to neuropsychology and are the journals most likely to be read by contributors to *Cortex*. Each, however, has its own niche or area of focus. Of the three journals, *Neuropsychologia* tends to be the journal for manuscripts on brain function in normal individuals. *Cortex*, as indicated by its mission statement, attracts papers about neuropsychological findings in patients with brain damage. *Brain* has a much wider range of articles, though it tends not to publish papers about brain function in normal individuals.

TABLE 28 – JOURNALS CITED BY *CORTEX* CITING AUTHORS

Cited Journals
Total Cited Journals: 3769
Top 10

	# Records	# Instances	Cited Journal
1	1357	1357	*Neuropsychologia*
2	1296	1296	*Cortex*
3	1275	1275	*Brain*
4	645	645	*Neurology*
5	568	568	*J Neuroscience*
6	557	557	*P Natl Acad Sci USA*
7	544	544	*Nature*
8	531	531	*Science*
9	478	478	*J Neurol Nurosurg PS*
10	376	376	*Arch. Neurol-Chicago*

B. Computational Linguistics

A taxonomy based on the citing papers was generated, using factor analysis. Among the high frequency phrases, there are no new applications within the central discipline identified, or research and applications external to the central discipline identified. Of course, this is not surprising, given that most of the citing papers were themselves published in *Cortex*.

IV. Computational Linguistics

Taxonomy Generation: Document Clustering

The 481 Cortex articles with Abstracts were clustered by the CLUTO algorithm into 32 elemental groups, yielding a high resolution average of fifteen records per group. These elemental groups were aggregated into different hierarchical levels. In the highest level, the 481 records were divided thematically into two categories; in the next highest level, the 481 records were divided into four categories; and so on. In the following analysis, the first three levels will be analyzed, and the themes (and associated numbers of records) of each category will be presented and discussed. The numbers in parentheses after the themes are the numbers of records.

CLUTO divides Level 1 into two categories: Handedness/ Awareness (211) and Memory (270).

CLUTO divides Level 2 into four categories, by dividing each Level 1 category into two sub-categories. Handedness/ Awareness (211) is divided into Handedness (145) and Neglect (66). Memory (270) is divided into Semantic Memory (151) and Amnesia (119).

CLUTO divides Level 3 into eight categories, by dividing each Level 2 category into two sub-categories. Handedness (145) is divided into Lateral Classification (82) and Lateral Movement (63). Neglect (66) is divided into Visual Field Neglect (38) and Neglect Diagnostics (28). Semantic Memory (151) is divided into Verbal/ Numerical (76) and Visual/ Spatial (75). Amnesia (119) is divided into Amnesia Symptoms (50) and Amnesia Physiology (69).

The following Cortex flat taxonomy can be generated. The bullets under each category represent the 32 elemental cluster themes.

HANDEDNESS (145)
-Lateral Classification (82)

- (13) - Selective attention.
- (12) - Ear asymmetry, especially in lateral discrimination with dichotic stimuli.

- (14) - Childhood dyslexia, especially association with deficits in inter-hemispheric interactions, as well as visual and language deficits.
- (12) - Immune and familial genetic disorders, emphasizing relation to laterality and handedness.
- (16) - Handedness experiments, and relation of handedness to other variables.
- (15) - Hand preferences, and the relationship of asymmetries to skills.

-Lateral Movement (63)
- (15) - Hand movements, especially manual asymmetries, for diagnosing apraxia effects.
- (11) - Handedness, especially in relation to motor functions, such as turning direction, reaching, grasping, both intra- and inter-manual..
- 0--(10) - Threshold detection, especially for hearing sounds, with some associations to simultaneous stimuli and bimanual tasks.
- (13) - Emotional stimuli, and hemispheric arousal related to facial expressions.
- 22--(14) - Hemispheric response differences to mainly visual stimuli, including color.

NEGLECT (66)
-Visual Field Neglect (38)
- (15) - Visual field stimuli, including dots and letters, emphasizing lateral imagery experiments.
- (11) - Extinction, emphasizing tactile but including other sensory inputs, and neglect, using contra-lesional and ipsi-lesional data.
- (12) - Neglect, in brain damaged patients, emphasizing right brain damage,

-Neglect Diagnostics (28)
- (13) - Line bisection tests, for evaluating neglect.
- (15) - Neglect, including personal and extra-personal, emphasizing cancellation experiments and left neglect.

SEMANTIC MEMORY (151)
-Verbal/ Numerical (76)
- (16) - Arithmetic and numerical calculations and facts, including trans-coding tasks..
- (17) - Priming, emphasizing word/ semantic, and its use in memory tests for AD patients.
- (22) - Reading and semantic/ word processing tests, especially for AD patients.
- (21) - Writing and word comprehension, primarily, for aphasic subjects, and speech and spelling/ grammatical errors secondarily.

-Visual/ Spatial (75)
- (17) - Name and word retrieval, especially of people and objects.
- (12) - Semantic categorization for living and non-living objects.
- (14) - Confabulation, especially in semantic and episodic memory tasks.

- (17) - Agnosia, primarily visual and tactile object recognition and naming disorders.
- (15) - Face recognition, emphasizing prosopagnosia, including overt and covert processing of facial identity.

AMNESIA (119)
-Amnesia Symptoms (50)

- (21) - Retrograde amnesia, with some related emphasis on anterograde amnesia and autobiographical memory.
- (14) - Learning, especially in Korsakoff's amnesia patients with memory problems.
- (15) - Amnesia, emphasizing forgetting rates, delays, and recall rates.

-Amnesia Physiology (69)

- (18) - Temporal lobe problems, especially in patients with temporal lobe epilepsy and/ or lobectomy, emphasizing memory impacts, and including hippocampel dysfunction as well.
- (17) - Script generation, mainly in patients with frontal lobe lesions, and associated executive function problems such as sequencing of actions and events for planning towards a goal.
- (17) - Memory, emphasizing short term studies, but including long term as well.
- (17) - Spatial orientation/ location problems, especially in relation to topographical memory, and the link to gyrus lesions, using associated regional cerebral blood flow and MRI measurements.

IV. SUMMARY AND CONCLUSIONS:

Publication Bibliomtrics

The *Cortex* top performers are centered in Italy primarily, and Western Europe generally. The Neuropsychologia top performers are centered in the UK primarily, and in the countries of the TTCP (The Technical Cooperation Program) totally. Recent text mining studies in the high tech areas of nanotechnology, nonlinear dynamics, fractals, among others, tend to have a much higher representation from the USA in top performers, some of the Asian countries like Japan, China, and South Korea, and Germany. Except for the USA, none of the top performers from these other countries are represented in these two neuropsychology journals.

It is somewhat surprising that such leading Canadian universities as McGill and the University of Toronto, British universities such as as Oxford and Cambridge, and American universities such as Harvard or the University of California system, are not represented in the *Cortex* list of prolific institutions. On the surface, the organizational country distributions appear to be cliquish in nature, with *Cortex's* prolific institutions being centered around the romance language countries, and *Neuropsychologia's* being centered around the English-speaking TTCP countries. Both journals might benefit from increased diversification.

Citation Bibliometrics

Even though *Cortex* has reasonable representation from the USA and UK in terms of numbers of authors, they are not getting similar representation in terms of numbers of highly cited papers from these two countries.

These most cited *Cortex* reference documents cover a much longer time domain than is typically found in text mining studies of physical science disciplines. For example, a recent study by the first author on nanotechnology examined a databsase of articles (from all journals) published in 2003. Of the twenty most highly cited articles, the oldest was 1991, and the median was 1996.

Researchers outside neuropsychology might wonder why more recent papers were not among the list of most cited works in the present study databases. In other fields, the older papers may have only historical interest because the field has moved far beyond the data and theories of that time. Neuropsychology appears to be different in that regard; change is slow and a method has to be useful and used for many years before it becomes a standard. Owing to the small numbers of patients with interesting syndromes, it is also likely that researchers keep a method long after its initial conception in order not to lose comparison between previous and former/future subjects in experiments.

The most cited articles in *Neuropsychologia* are cited, on average, more than three times as often as the most cited articles in *Cortex*, and the most cited articles in *Brain* are cited,

on average, more than twice as often as the most cited articles in *Neuropsychologia*. Whether the difference in highly cited papers is due to the difference in intrinsic quality of the best papers in each journal, the thrust areas selected within the neuropsychology discipline, or the number of people who have access to each journal, or some combination of these causes, cannot be stated at this time.

The average number of citations increases with the average number of authors, ranging from about four authors of the most cited *Cortex* papers to about seven authors of the most cited *Brain* papers. Having more authors may add more dimensions and perspectives to a paper, increasing its comprehensiveness, and having more authors may be the equivalent to having more peer reviewers, increasing the paper's quality.

The most cited papers have substantially more references than the least cited, in both journals, and the effect is most pronounced in *Neuropsychologia*. Additionally, the average number of citations increases with the average number of references, ranging from about 46 references in the most cited *Cortex* papers to about 68 references in the most cited *Brain* papers. Having more references is one measure of increased scholarship, and may result in additional citations, on average, for historical purposes.

The most significant country representations in the samples examined are the strong positive showing of the UK, and the weak showing of Japan. This latter observation is very different from Japan's strong showings in almost all of the high tech discipline text mining studies performed by the first author, and suggests that Japanese scientists have not concentrated their energies in the area of neuropsychology.

There is a distinct shift in type of study in proceeding from *Cortex* to *Neuropsychologia* to *Brain*. Clinical behavioral studies, many of them essentially case studies, predominate the most cited *Cortex* papers. *Neuropsychologia* has more of a balance between Behavioral and Diagnostic-Non-Invasive (e.g., PET, MRI) in its ten most cited papers. Brain shows a heavy emphasis on Diagnostic-Non-Invasive (7/10), two papers on surgical procedures, and one on Diagnostic-Invasive (e.g., tissue samples). Based on reading Abstracts from each of these journals, the types as represented in the top ten most cited articles roughly approximate the types of papers published overall. Thus, as citations increase in absolute amounts, the study type transitions from the clinically oriented behavioral focus to the correlates with more objective measurements.

Finally, these bibliometrics results suggest that *Cortex* needs to take steps to attract a more diverse group of highly prolific and cited contributors than its continental Western European base, if it wishes to capture a greater share of the seminal neuropsychology papers. Further investigation of these citation differences is recommended.

Citation Mining

Papers published in *Cortex* and *Neuropsychologia* are of fundamental importance to subsequent papers to be published in each of the journals. To determine whether this has its origin in a "niche" occupied by the journals or instead implies a degree of narrowness

would require a different kind of analysis, one requiring a close analysis of the contents of the papers citing the *Cortex* and *Neuropsychologia* articles. The results of the journal citation comparison study suggest that for both journals many of the citations are intra-journal; in the case of *Neuropsychologia*, the next-cited journal is cited less than half as many times as *Neuropsychologia* itself. Further analyses are required to determine the reason(s) for this.

Taxonomy

Finally, document clustering showed that the articles sampled in *Cortex* can be reasonably divided into four categories (papers in each category in parenthesis): Semantic Memory (151); Handedness (145); Amnesia (119); and Neglect (66). A similar cluster analysis has not been carried out for *Neuropsychologia* or *Brain*, but the general categories are probably similar, though not necessarily in the same order of importance.

REFERENCES

1. Kostoff, R. N., "Database Tomography for Technical Intelligence," Competitive Intelligence Review, 4:1, Spring 1993.

2. Kostoff, R.N., "Database Tomography: Origins and Applications," Competitive Intelligence Review, Special Issue on Technology, 5:1, Spring 1994.

3. Kostoff, R. N., Eberhart, H. J., and Miles, D., "System and Method for Database Tomography", U. S. Patent Number 5440481, August 8, 1995.

4. Kostoff, R. N., "Database Tomography for Technical Intelligence: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society:, Scientometrics, 40:1, 1997.

5. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature". Information Processing and Management. 34:1. 1998.

6. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography". Journal of the American Society for Information Science. 15 April 1999.

7. Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. Jan-Feb 2000.

8. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". Journal of Aircraft, 37:4. 727-730. July-August 2000.

9. Kostoff, R. N., and Geisler, E. "Strategic Management and Implementation of Textual Data Mining in Government Organizations". Technology Analysis and Strategic Management. 11:4. 1999.

10. Losiewicz, P., Oard, D., and Kostoff, R. N. "Textual Data Mining to Support Science and Technology Management". Journal of Intelligent Information Systems. 15. 99-119. 2000.

11. Kostoff, R. N., and DeMarco, R. A. “Science and Technology Text Mining”. Analytical Chemistry. 73:13. 370-378A. 1 July 2001.

12. Kostoff, R. N.. “Text Mining for Global Technology Watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799. 2003.

13. Hearst, M.A.. "Untangling Text Data Mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.

14. Zhu, D.H., and Porter, A.L.. "Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting". Technological Forecasting and Social Change. 2002. 69 (5): 495-506.

15. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", Journal of Information Science, 23:4, 1997.

16. Science Citation Index. ISI Web of Science. Thomson ISI. Phila., PA.

17. TechOasis. Search Technology, Inc. Norcross, GA.

18. Kostoff, R. N. "Science and Technology Innovation". Technovation. 19:10. 593-604. October 1999.

19. Kostoff, R. N. "Stimulating Innovation". International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 2003.

20. Swanson, D.R., "Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge", Perspect Biol Med, .30: (1), 1986.

21. Smalheiser, N.R., Swanson, D.R., "Using ARROWSMITH: A Computer Assisted Approach to Formulating and Assessing Scientific Hypotheses", Comput Meth Prog Bio, 57: (3), 1998.

22. Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. "Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling". JASIST. 52:13. 1148-1156. 52:13. November 2001.

23. WINSTAT. R. Fitch Software. St.-Martin-Allee 1. D-79219 Staufen, Germany.

24. Lotka, A. J. (1926) The Frequency Distribution of Scientific Productivity. Journal of the Washington Academy of Sciences. 16.

25. MacRoberts M, MacRoberts B. Problems of citation analysis. Scientometrics 1996; 36(3): 435-444.

26. Kostoff, R. N. "The Use and Misuse of Citation Analysis in Research Evaluation". Scientometrics. 43:1. September 1998.

27. Carey, D.P. Citation impact of individuals and journals. Cortex Forum on Impact Factor, Cortex. 37:4. 580-582. 2001.

28. Axelrod, B.N and Millis, S.R. Preliminary standardization of the Cognitive Estimation Test. *Assessment*, 1:269-274, 1994.

APPENDICES

APPENDIX A. TAXONOMY GENERATION: Non-Statistical Clustering

1. Abstract Field Database: 1991-2001

Phrase frequencies were generated for 10 years of *Cortex* Abstracts (1991-2001). High technical content phrases were identified, and then classified into a three-level taxonomy. The resulting taxonomy and a sample of related phrases is given in Table A-1.

Table A-1

NON	MEMORY	DISORDER Alzheimer's Disease Amnesia Retrograde Amnesia
		GENERAL Semantic Memory Episodic Memory Working Memory
		TEST Serial Position Curve Memory Tests
	ASSOCIATION	DISORDER Double Dissociation
		GENERAL Face Recognition Object Recognition
		TEST Familiar Faces Picture Naming Object Naming
	VISUAL	DISORDER Visual Agnosia Optic Aphasia
		GENERAL Visual Field Visual Processing Visual Stimuli
		TEST Line Bisection
	AUDITORY	GENERAL Left Ear Right Ear

		Ear Advantage
	LINGUISTIC	DISORDER Aphasia
		GENERAL Angular Gyrus
		TEST Verbal Fluency Lexical Decision
	GENERAL	TEST Decision Task
MOTOR	MOTOR	DISORDER Parkinson's Disease Parkinson's Disease Patients
		GENERAL Hand Hand Preference Basal Ganglia
		TEST Finger Tapping
GENERAL	GENERAL	GENERAL Temporal Lobe Left Hemisphere Normal Controls
		TEST Reaction Time Task Performance
	DISORDER	GENERAL Agnosia Brain Damage Hemisphere Lesions
		NEGLECT Neglect Personal Neglect

Based on the contents of this taxonomy, the following underlying characteristics of Cortex can be inferred.

The journal Cortex covers a wide variety of topics, most of which can arbitrarily be divided into Motor and Non-Motor emphasis areas. A third category, GENERAL, covers other topics unassociated with these two categories (although logically, one would have thought that the first two emphasis areas would cover all topics. The Non-Motor areas predominate, with a strong emphasis (based upon phrase frequencies for category related phrases) on MEMORY, VISUAL, LINGUISTIC and ASSOCIATION related topics, and secondary emphasis on NEGLECT and AUDITORY. Much of the focus centers around the cognition/ observation aspects and in clinical settings, with emphasis on patients.

According to this analysis, there appears to be very little mention of the theoretical work, little or no mathematical modeling, and virtually no discernable direct theoretical links to biology at the genetic and molecular level.

2. Full Text, Abstract, Title, Keyword Fields Databases: 1997-2001

Phrase frequencies were generated for four years of Keywords, Titles, Abstracts, and Full Text of Cortex articles from 1997 to 2000. The four databases then served as one basis for comparison among the four fields. Table A-2 contains some metrics for comparing phrase frequencies among databases.

The first metric, SINGLE WORD (the column headed by SINGLE WORD), represents the total number of single word phrases in each database (no cut-off frequency). The second and third metrics, DOUBLE WORD and TRIPLE WORD, have analogous meanings for double and triple word phrases. The fourth metric, COMBINATION, is a sum of the single, double, and triple word phrase totals. The first column represents the databases in which the phrases appeared. For example, the 10 phrases in the K ONLY - SINGLE WORD PHRASES matrix element appeared only in the Keywords, and not in the Abstract, Title, or Full Text. Fields with two or more letters represent phrases that appeared in a combination of databases (e.g. phrases in the TK category appeared in both the Title and Keyword databases, but not the Abstract or Full Text databases.) The TOTAL PHRASES field at the bottom lists the total number of phrases in each of the databases (e.g., there are 676 Keyword phrases total.)

A sampling across frequency bands in the larger text fields showed that about 1/3 of the phrases could be classified as high technical content. This number was relatively invariant to frequency. Thus, the Abstract field contains about an order of magnitude more technical phrases than the Title or Keyword fields, and the Full Text contains more than an order of magnitude more phrases than the Abstract.

One problem with the Keyword field should be noted. The SCI has two fields for Keywords (Author Keywords and Keywords Plus), but they were combined for analytical purposes in this paper. Keywords plus appears to be third-party indexer generated, and random checks showed there could be substantial disparities between the emphasis areas of the Abstract and those reflected by Keywords Plus. Thus, even the modest numbers reported for Keywords in Table A-2 should be reduced to reflect the mismatch between Keywords Plus and areas of emphasis.

Table A-2 – Unique Phrases in each Field and Field Combination

Data-Base	Single Word	Double Word	Triple Word	Combination
K only	10	63	32	105
T only	7	152	369	528
TK	1	0	0	1
A only	128	2361	5301	7790
AK	1	0	2	3
AT	2	11	25	38
ATK	0	0	0	0
F only	19657	95971	152178	267806
FK	53	83	29	165
FT	77	164	129	370
FTK	5	6	4	15
FA	2749	3293	2288	8330
FAK	125	65	16	206
FAT	520	250	104	874
FATK	117	57	7	181
SUM	23452	102476	160484	286412

TOTAL PHRASES	
Key-word	676
Title	2007
Abstract	17422
Full Text	277947

LEGEND

K = KEYWORD

T = TITLE

A = ABSTRACT

F = FULLTEXT

Figure A-1 is a plot of the distribution functions for the frequencies of unique phrases in each database. The ordinate represents the number of phrases f(N) with frequency N, and the abscissa represents the frequency N.

FIGURE A-1

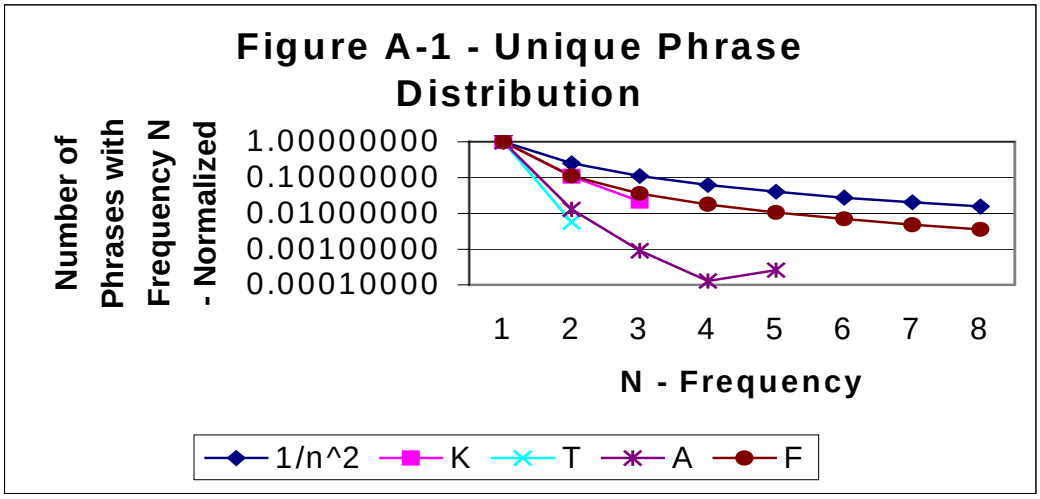

Table A-3 lists the 20 highest frequency high technical content phrases for each of the four databases. All four databases focus on impairment/ disease, memory, visual, and neglect. The highest technical content group is the Keywords, since they reflect summary descriptions, while the lowest technical content group in this highest technical content range is Full Text.

TABLE A-3 – HIGHEST FREQUENCY HIGH TECHNICAL CONTENT PHRASES

KEYWORD		TITLE		ABSTRACT		FULLTEXT	
FREQ	PHRASE	FREQ	PHRASE	FREQ	PHRASE	FREQ	PHRASE
25	Impairment	25	Memory	269	Patients	4106	Patients
22	Deficits	25	Patients	155	Left	2748	Memory
20	Retrieval	23	Neglect	148	Memory	2481	Subjects
18	Dementia	18	Disease	132	Right	2384	Left
18	Lesions	17	Case	118	Visual	2336	Two
17	Memory	17	Study	110	Tasks	2218	Task
16	Attention	16	Semantic	109	Neglect	2198	Visual
14	Performance	16	Visual	107	Performance	2135	Right
13	Alzheimer's Disease	15	Alzheimer	107	Subjects	2075	Test
13	Aphasia	14	Patient	103	Two	1989	Performance
12	Perception	14	Right	99	Semantic	1878	One
12	Visual neglect	13	Evidence	93	Task	1797	Semantic
11	Damage	13	Left	89	Patient	1735	Patient
11	Knowledge	12	Amnesia	89	Showed	1681	Tasks
11	Patient	9	Dissociation	87	Normal	1636	Group
11	Recognition	9	Injury	77	Results	1491	Words
11	Semantic memory	9	Processing	67	Hand	1398	Neglect

10	Brain	8	Brain	64	Processing	1336	Study
10	Dissociation	8	Frontal	64	Temporal	1333	Errors
10	Language	8	Naming	58	Control	1305	Significant

Table A-4 lists phrases in each database that are unique[1]; i.e., not contained in any of the other databases. The stems of these words are also absent from the other fields. [Footnote: 1: The ABSTRACT phrases are not located in the Keywords or Titles, but they may or may not be in the Full Text]

The first Keyword is an example of the Keywords Plus problem mentioned previously. It occurred in one article, whose subject was lateral bias in the prehensile tail use of monkeys. The Author Keywords were LATERALITY, PREHENSILE TAIL, and SPIDER MONKEY, while the Keywords Plus expanded the list to: LEMUR LEMUR-CATTA, CEBUS-APELLA, MANUAL LATERALITY, MULTIPLE MEASURES, SQUIRREL-MONKEYS, HAND PREFERENCE, ASYMMETRIES, HANDEDNESS, BRAIN, and RAT. The Author Keywords were presumably chosen by the author(s) to reflect the main themes of the article, as detailed in the Abstract, while the Keywords Plus represent minor themes, for the most part, which may or may not be useful to potential readers in guiding them to this particular paper.

Obviously, the many tens of thousands of unique Full Text phrases could not be listed here, but the few listed shows they are not trivial. That is, they would appear to be valuable when populating thematic categories with detailed sub-themes. Anosognosia, the first word in the Abstract field, for instance, may be present in its "deficit" form in the full text (anosoagnosia – denial of one's own illness), or in the form found in the full text, anosodiaphoria. Therefore, if the search database looks only at key words and abstracts, a search for anosoagnosia or anosodiaphoria would miss this article, It is easy to understand why terms are found in the full text but not in the list of key words or in the abstract, but it's surprising that there are terms in these sections that are not found again in the full text.

TABLE A-4 – UNIQUE PHRASES IN EACH DATABASE

KEYWORD	ABSTRACT	FULLTEXT
Cebus Apella	Anosognosia	Cornea
Hemidecortication	Ambidextrous	Cataracts
Hypercalcemia	Braille	Carotid
Oppel-Kundt	Tachistoscopic	Cancer
	Pontine	Calcarine
	Paroxystic	Antibiotic
	Cerebral Perfusion	Anosodiaphoria
	Autosomal Dominant	Alloesthesia
	Operculum	Ahylognosia
	Olfactory	Angiography
		coronal

adenoma
anagram
caloric
apathy
arcuate fasciculus
autonoetic consciousness
chromosome

The taxonomies generated for the four fields were almost identical to that created for the ten-year Abstract database. However, a few differences emerged. The Title field showed the biggest deviation. It almost completely lacks any testing- related phrases. The disorders and non-disorders are present however. The Keyword field not only lacks an Auditory category, but also did not have the detail seen in the other fields. The four-year Abstract database produced the same taxonomy as the ten-year Abstract database, but with less detail, as was expected. The Full Text field had an enormous amount of detail in each category, providing almost innumerable examples for each category.

Why are these differences important? All levels of text mining, ranging from standard information retrieval to the more exotic literature-based discovery, tend to access records through phrase matching. As the results show, there can be substantial differences in records retrieved, depending on which fields are accessed by the search engines. For high-level taxonomy generation, the field differences are less severe, but when lower level taxonomic detail is required, then the differences become important. For literature-based discovery in particular, the predominant publishing group (17, 18) has used Title and Keyword phrases for information processing almost exclusively, and it is obvious there is much literature content not being accessed if this restriction is applied.

APPENDIX B. TAXONOMY GENERATION: Statistical Clustering

For each square matrix of high frequency phrases used to generate the higher level taxonomy, two analyses were run: multi-link statistical clustering, and factor analysis.

1. Multi-Link Clustering

The multi-link statistical clustering resulted in three types of raw data output:

1) A dendogram that shows the quantitative linkages among closely-related phrases. Figure B-1, for example, is a dendogram that portrays linkages among the twenty highest frequency technical content phrases from the Abstracts database. The abscissa represents each phrase, and the ordinate is a distance metric of the 'closeness' of each phrase. 'Closeness' of two phrases is the similarity in their profile of co-occurrence with the other phrases. One key advantage of the dendogram is that not only are the clusters easily portrayed, but one can easily see the position of each phrase relative to the other phrases, which is an important feature of the inter-relations of the phrases. Sometimes unique insights emerge when attempting to explain seemingly anomalous positionings of phrases.

2) A table that contains a quantitative measure of the similarity of adjoining phrases or phrase-cluster pairs. The similarity, or 'distance', is obtained by matching the co-occurrence profiles. Table B-1, for example, contains the information portrayed in Figure B-1. This table is valuable for studying the sequencing of cluster development. Steps 1-4 represent the creation of four individual units from eight separate phrases. Step 5 represents the combination of one of the units WORDS (consisting of the phrases words and reading, as shown in Step 2) with the phrase WRITING to form a higher level unit.

3) A taxonomy of a pre-specified number of groups of phrases. Table B-2, for example, shows the groupings of phrases when four clusters were specified for the data portrayed in Figure B-1.

TABLE B-1 – HIERARCHICAL CLUSTER DEVELOPMENT

Step	joining Cluster 1	Size 1	with Cluster 2	Size 2	Distance
1	left hemisphere	1	right hemisphere	1	31.39987248
2	words	1	reading	1	34.59715688
3	handedness	1	hand	1	36.46334392
4	impairment	1	deficit	1	37.5820787
5	words	2	writing	1	38.02428564
6	normal	1	naming	1	38.51191888
7	objects	1	amnesia	1	38.99446727
8	words	3	normal	2	39.06481799
9	memory	1	recall	1	39.06609507
10	impairment	2	objects	2	39.28078964
11	left hemisphere	2	language	1	39.43335545
12	normal subjects	1	recognition	1	39.49334066
13	impairment	4	memory	2	39.79451751
14	normal controls	1	left hemisphere	3	39.90021232
15	neglect	1	normal subjects	2	40.0532002
16	impairment	6	neglect	3	40.11166017
17	impairment	9	words	5	40.2376026
18	impairment	14	normal controls	4	40.27576376
19	impairment	18	handedness	2	40.30990082

FIGURE B-1

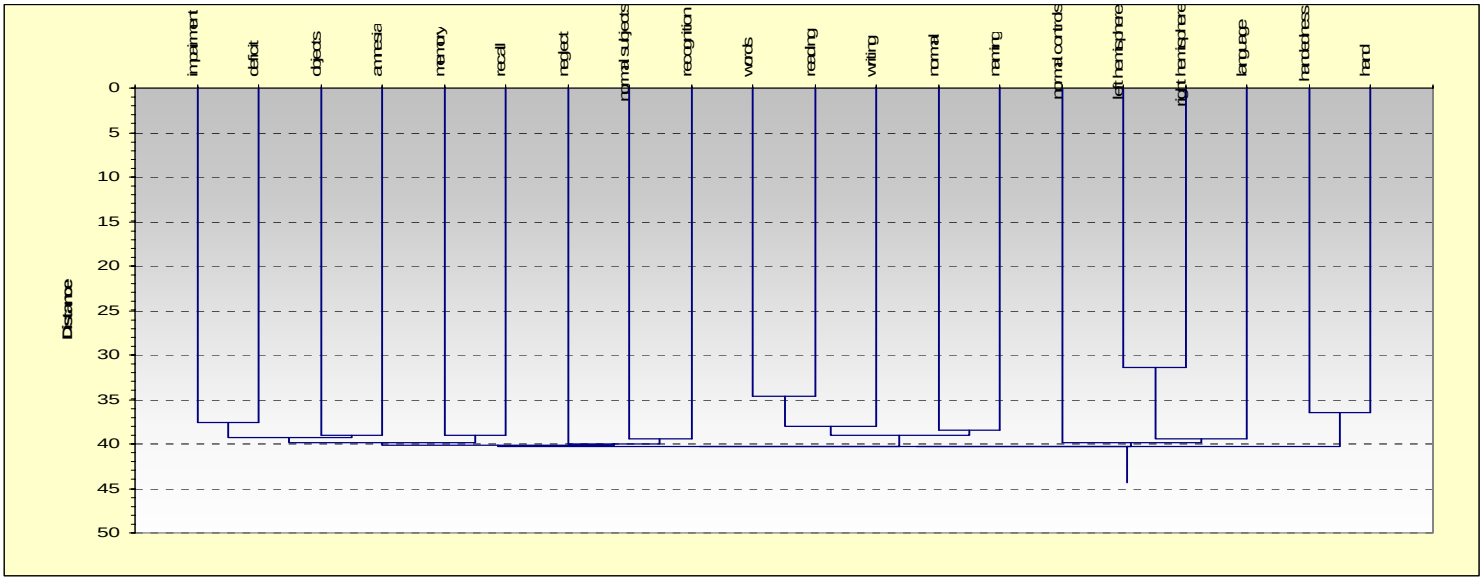

TABLE B-2 – FOUR CLUSTER TAXONOMY

CLUSTER	PHRASE
1	impairment
1	deficit
1	memory
1	objects
1	neglect
1	normal subjects
1	recognition
1	recall
1	amnesia
2	words
2	normal
2	naming
2	reading
2	writing
3	handedness
3	hand
4	normal controls
4	left hemisphere
4	right hemisphere
4	language

2. Factor Analysis

The second statistical clustering technique used is Factor Analysis. It generates a specified number of factors, each of which represents a focal area in the journal articles. Table B-3 is a sample output of a factor analysis with 4 factors chosen, using the same phrases as in the above clustering run. The numbers under the factor headings are factor loadings, or correlations between the phrase and the factor. Those loadings that are above 0.25 are the darkest highlighted, and those between 0.10 and 0.20 are the lightest highlighted.. The darkest phrases under each factor are the major components of the factor, and a suitable heading can be derived by looking at these phrases. Thus, Factor 1 would be a MEMORY or AMNESIA category, Factor 2 would be HEMISPHERE LATERALIZATION, Factor 3 would be LINGUISTIC, possibly focusing on written aspects, and Factor 4 would be MOTOR or MOTOR LATERALIZATION. Another useful aspect of the factor analysis is the ability to determine which phrases are multidisciplinary. The phrase WRITING has a strong loading on factor three, the linguistic category, and a smaller but still substantial loading on factor four, the motor category. Such linkings are useful for relating various factors to one another. The communality column is the sum of squares of the factor loadings in each row, and shows each phrase's overall loading on all the factors.

TABLE B-3 – FOUR FACTOR TAXONOMY

Phrase	Factor 1	Factor 2	Factor 3	Factor 4	Communality
amnesia	0.27	0.00	0.01	0.03	0.070971034
objects	0.26	0.01	0.00	0.02	0.068056413
memory	0.25	0.00	0.03	0.02	0.063543078
impairment	0.25	-0.02	0.04	-0.01	0.064022057
recall	0.25	0.00	0.03	0.02	0.063448781
deficit	0.24	-0.02	0.07	-0.03	0.061060124
recognition	0.23	0.04	0.02	0.06	0.060585822
neglect	0.23	0.04	0.04	0.08	0.062436895
normal subjects	0.23	0.03	0.04	0.09	0.062227208
normal controls	0.22	0.04	0.04	0.07	0.057416942
language	0.21	0.08	0.03	0.07	0.056145766
normal	0.19	0.00	0.12	0.04	0.052990472
right hemisphere	0.06	0.52	-0.03	0.01	0.273660234
left hemisphere	0.03	0.52	0.05	0.01	0.272936036
reading	-0.02	0.02	0.47	0.00	0.218609519
words	0.07	-0.01	0.36	-0.01	0.132339476
naming	0.14	0.00	0.23	0.01	0.073542117
writing	0.06	0.03	0.23	0.16	0.082282181
handedness	0.07	0.00	0.02	0.42	0.183339121
hand	0.09	0.00	0.01	0.39	0.161419955

3. Abstract Field Database: 1991-2001

Table B-4 shows the final taxonomy from the results of the multi-linked clustering algorithm. Four overarching categories were generated, based on the dendogram and personal understanding of the overall discipline. They are, in no specific order, VISUAL, LINGUISTIC, MEMORY, and a combined MOTOR / LATERALIZATION / NEGLECT category.

Table B-4 – Multi-Link Clustering Taxonomy

Visual	Linguistic	Memory			Motor / Lateralization / Neglect	
Internal Manipulation of Object Images	Written Language Processing	Phonology/ Short Term Memory Deficiency	General Memory & Loss (Both STM & LTM)	Memory & Loss	Cerebral Dominance	Oral Communication (Both Verbal & Visual Aspects)
Hemisphere Disconnect Based Visual Dissociation	Oral Language Processing	Episodic/ Semantic Memory Synergy	Verbal-Based Color Memory & Recognition	Motor Memory	Personal Neglect & Related Functions	Spatial Lateral Bias
Lateral Visual Tasks		Name Anomia	Learning & Deficits	Tactile Memory	Writing Motor Deficiencies	Motor Control Deficiencies (Neglect)
Visual Processing and Defects		Acalculia	Illiteracy	AD / Dementia & Testing	Language Lateralization	
Hemisphere-Disconnection-Based Visual Ordering		Organically-Based Spatial Amnesia	Ordering Deficit	Face/ Object Recognition Impairments		

Table B-5 shows the final taxonomy from the results of the factor analysis. Table B-5 is similar in structure to Table B-4, but contains 248 high technical content phrases. For display purposes only, any phrase with less than 0.30 loading on at least one factor was omitted.

Parametric runs were made ranging from 2 to 10 factors, and the seven factor result was judged to be the most realistic. The factors are, in order: CEREBRAL DOMINANCE &

MOTOR LATERALIZATION; VISUAL ASSOCIATION; LINGUISTICS; NEGLECT; LEARNING; PROSOPAGNOSIA; and NAME ANOMIA. These factors/ categories can be combined into the same higher level categories generated from the multi-linked clustering algorithm: visual, linguistic, memory, motor / lateralization / neglect.

TABLE B-5 – SEVEN FACTOR TAXONOMY

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Communality
handedness	0.69	-0.09	-0.02	0.00	-0.03	-0.02	0.00	0.481404234
hand preference	0.62	-0.08	-0.03	0.01	-0.01	-0.03	-0.01	0.391422894
hand	0.58	-0.06	-0.03	-0.07	-0.14	0.00	-0.11	0.382715199
left-handers	0.56	-0.10	-0.06	-0.02	-0.01	0.04	0.04	0.33644657
right-handers	0.56	-0.10	-0.07	-0.02	-0.01	0.02	0.04	0.3354815
left hand	0.54	0.03	0.00	-0.06	-0.04	-0.02	-0.07	0.300045661
eyedness	0.51	-0.08	-0.04	-0.02	-0.01	0.00	0.03	0.267694541
right-handed subjects	0.49	-0.05	-0.07	0.02	-0.04	0.02	0.00	0.245967049
right hand	0.47	0.07	0.00	-0.06	0.01	-0.03	-0.07	0.2394355
finger tapping	0.44	-0.06	0.00	0.04	-0.03	-0.01	0.01	0.198113385
manual asymmetries	0.43	-0.05	-0.05	-0.02	-0.10	0.01	-0.10	0.211561768
left-handedness	0.41	-0.07	-0.03	-0.01	0.01	0.01	0.04	0.178027593
movements	0.41	-0.02	-0.05	-0.04	-0.11	-0.02	-0.12	0.199886617
cerebral dominance	0.39	-0.08	-0.03	-0.04	-0.06	0.00	-0.04	0.163171659
hand movements	0.38	-0.05	-0.09	-0.10	-0.12	0.00	-0.10	0.192519386
asymmetries	0.38	-0.03	0.01	-0.01	-0.08	-0.01	-0.01	0.152727625
objects	-0.14	0.70	-0.08	-0.10	0.03	-0.05	0.11	0.545426422
associative agnosia	-0.07	0.63	-0.03	-0.06	-0.05	0.00	0.17	0.446143123
line drawings	-0.06	0.54	-0.08	0.16	0.02	0.04	0.11	0.337124852
drawing	-0.07	0.53	-0.06	-0.02	0.06	0.01	0.05	0.299687715
drawings	-0.07	0.49	-0.05	0.24	0.04	0.03	-0.01	0.306360656
left posterior cerebral artery	-0.06	0.42	0.05	0.05	-0.02	-0.02	-0.07	0.189192503
agnosia	-0.02	0.41	-0.07	0.04	-0.02	0.04	0.25	0.237059902
visual stimuli	-0.08	0.39	-0.02	0.23	-0.05	-0.03	-0.08	0.218306361
optic aphasia	-0.07	0.38	0.02	-0.17	-0.11	-0.02	0.02	0.194839939
visual processing	-0.06	0.36	-0.05	-0.04	-0.02	0.10	-0.08	0.149843443
visual imagery	-0.08	0.35	-0.06	0.13	0.05	0.00	0.07	0.155191245
visual input	-0.07	0.34	-0.01	-0.05	-0.01	-0.02	0.01	0.124969295
pantomimes	-0.02	0.33	0.05	-0.16	-0.09	0.02	-0.04	0.15139819
photographs	-0.06	0.33	-0.07	0.06	0.05	0.14	0.18	0.176520506
visual agnosia	-0.03	0.32	-0.04	-0.13	0.04	-0.01	-0.11	0.138440876
object	-0.08	0.30	-0.15	-0.26	-0.07	0.00	0.07	0.19740965
shapes	0.07	0.30	-0.06	0.15	-0.03	0.00	0.05	0.122460926

dictation	-0.09	-0.08	0.72	-0.05	-0.01	-0.03	-0.08	0.541825251
reading	-0.08	-0.10	0.63	0.00	-0.05	-0.05	-0.07	0.42570269
writing	0.08	-0.05	0.62	-0.06	-0.05	-0.04	-0.04	0.401035391
oral spelling	-0.05	-0.09	0.61	-0.03	0.00	0.00	0.00	0.382349369
naming	-0.13	0.01	0.61	-0.11	-0.03	0.04	0.18	0.436640291
repetition	-0.07	-0.09	0.58	-0.04	-0.05	0.03	-0.01	0.354159999
semantic store	-0.09	-0.09	0.53	-0.06	0.00	0.03	0.23	0.351163887
spelling	-0.06	-0.05	0.51	0.01	-0.02	-0.05	-0.07	0.273652306
oral reading	-0.03	-0.07	0.43	-0.02	0.00	0.02	0.00	0.18894216
comprehension	-0.05	-0.08	0.42	-0.01	-0.03	-0.04	0.02	0.186443343
modality	-0.09	0.02	0.39	-0.07	0.01	-0.05	-0.07	0.172559773
selective deficit	-0.13	-0.02	0.38	-0.06	0.05	-0.05	-0.08	0.177790464
picture naming	-0.04	-0.03	0.36	-0.03	0.02	0.03	0.06	0.137608386
normal	-0.15	0.04	0.32	-0.16	0.12	0.06	-0.02	0.174362081
neglect	-0.08	0.01	-0.06	0.69	-0.08	0.00	-0.07	0.497618666
neglect syndrome	-0.07	0.07	-0.05	0.66	-0.06	0.00	-0.09	0.455431295
anosognosia	-0.02	0.00	-0.05	0.63	-0.06	0.02	-0.08	0.409676442
vestibular stimulation	-0.03	0.01	-0.04	0.60	-0.05	0.02	-0.07	0.366980962
personal neglect	-0.04	0.07	-0.04	0.58	-0.03	0.02	-0.06	0.350354477
neglect patients	-0.05	0.04	-0.05	0.52	-0.03	0.00	-0.06	0.285000785
motor performance	0.29	-0.01	-0.01	0.40	-0.06	0.00	-0.09	0.250432437
neglect dyslexia	-0.12	-0.02	-0.04	0.32	-0.10	-0.15	-0.03	0.153507093
left unilateral neglect	-0.03	0.22	-0.05	0.31	-0.01	0.01	-0.02	0.151252993
spatial neglect	-0.05	-0.05	-0.03	0.30	-0.01	-0.05	-0.01	0.096865134
retention	-0.07	-0.09	-0.03	-0.07	0.63	0.03	-0.07	0.419802552
verbal learning	-0.03	-0.16	-0.11	0.03	0.55	-0.07	-0.01	0.343521433
verbal memory	-0.03	-0.15	-0.09	0.02	0.53	-0.06	0.00	0.314176865
learning	-0.14	-0.16	-0.19	-0.08	0.52	-0.11	-0.08	0.376624117
right temporal lobe	-0.04	-0.17	-0.10	0.05	0.48	-0.03	0.00	0.270726621
encoding	-0.07	-0.14	-0.12	-0.03	0.45	-0.13	-0.06	0.264015468
impairment	-0.21	-0.06	0.04	-0.17	0.40	0.06	0.08	0.246652636
retrograde memory	-0.07	0.04	0.07	-0.10	0.38	0.13	-0.04	0.187076325
severe anterograde amnesia	-0.05	0.01	0.03	-0.05	0.36	0.12	0.00	0.152707057
frontal lobes	-0.03	-0.11	-0.11	-0.03	0.34	-0.16	-0.02	0.169406138
side	0.01	-0.08	-0.04	0.19	0.33	-0.07	-0.03	0.162488836
autobiographical events	-0.09	0.06	0.04	-0.11	0.33	0.09	-0.05	0.146812641
hippocampus	-0.13	-0.09	-0.12	-0.12	0.31	-0.14	-0.15	0.186637953
face processing	-0.14	-0.15	-0.13	-0.08	-0.13	0.66	-0.07	0.525363568
familiar faces	-0.13	-0.14	-0.09	-0.08	-0.07	0.64	-0.09	0.477186464
prosopagnosia	-0.11	-0.18	-0.06	0.05	0.05	0.63	0.13	0.460692713
unfamiliar faces	-0.10	-0.18	-0.06	0.03	-0.01	0.62	0.09	0.433962279
traumatic brain injury	-0.10	-0.17	-0.10	-0.02	0.00	0.52	-0.06	0.318827538
face	-0.13	-0.01	-0.03	-0.11	0.15	0.45	0.02	0.257080433

impairments	-0.08	-0.15	0.01	-0.01	-0.10	0.38	-0.05	0.186104165
object agnosia	-0.01	-0.12	-0.02	0.12	0.03	0.35	0.22	0.201871957
breakdown	-0.13	-0.08	-0.14	-0.09	-0.12	0.34	-0.14	0.206885369
face recognition	-0.10	-0.03	-0.12	-0.13	-0.19	0.32	-0.09	0.188129197
names	-0.16	-0.14	0.01	-0.09	-0.07	0.09	0.69	0.545810026
bilateral lesions	-0.05	0.18	-0.11	-0.13	-0.05	-0.03	0.58	0.401555058
temporal lobes	-0.07	0.02	-0.09	-0.08	-0.07	-0.10	0.48	0.268526712
common names	-0.10	-0.12	0.10	-0.05	-0.06	0.02	0.47	0.263047314
people's names	-0.09	-0.14	-0.02	-0.06	0.01	-0.03	0.43	0.21833193
left temporal lobe	-0.04	-0.03	-0.02	-0.06	-0.02	-0.02	0.39	0.162137384
faces	-0.10	0.15	-0.04	-0.08	-0.07	0.20	0.39	0.23844803
parietal lobes	-0.04	0.11	-0.08	-0.13	-0.07	-0.02	0.39	0.193347403
spatial tasks	-0.04	0.06	-0.08	-0.08	-0.03	-0.04	0.36	0.154472456
anterograde amnesia	-0.09	-0.05	-0.12	-0.14	-0.03	-0.06	0.35	0.170006598
proper name anemia	-0.08	-0.13	0.04	-0.02	-0.05	-0.03	0.34	0.143889888
name	-0.14	-0.09	-0.01	-0.07	-0.12	-0.11	0.32	0.166000762
semantic information	-0.17	0.01	-0.06	-0.10	-0.03	-0.11	0.32	0.157592216

4. Full Text, Abstract, Title, Keyword Fields Databases: 1997-2000

The same techniques were performed on a shortened database spanning the articles from early 1997 through 2000. Four different fields were analyzed, the Titles, Keywords, Abstracts, and the Full text. However, the Title field had insufficient phrases for the factor analysis to generate a meaningful categorization. Taxonomies for the remaining fields were generated using each of the two statistical techniques.

The factor analysis proved more useful in generating the taxonomies, and showed a significant difference between the three analyzed fields.

Abstract: motor lateralization; name anomia; memory; callosal disconnection; prosopagnosia; oral linguistics; and neglect. The absence of a visual association category could be a result of a change of the area of focus of research in the time periods studied.

Full Text (with example phrases from each category): memory and loss of memory (MEMORIES and AMNESIA); visual association and semantic representations (OBJECTS, NAMING, and PHOTOGRAPHS); lateralized neglect being tested with visual stimuli (NEGLECT, ANOSOGNOSIA, and VISUAL FIELD); motor systems (arms/ hands/ fingers only) and its lateralization (LEFT HAND, MOVEMENTS, and AGRAPHIA); oral communication (APHASIA, SPEECH, and LANGUAGE); Alzheimer's disease and dementia (DEMENTIA and ALZHEIMER'S DISEASE); physical brain damage and imaging (BRAIN LESION, HEAD TRAUMA, and PET/ MRI/ CT); and written linguistics and mathematics (WRITING, SUBTRACTION, and READING).

Keyword: neglect; acalculia; lateralization of the motor system; memory; visual attention; visual reading; general deficits; apraxia; and visual imagery.

The Abstract field generated was nearly identical to that generated in the 10 year database, with various categories dealing with both generalized and specific topics. The Full Text taxonomy created categories that mirrored not only the Abstract taxonomy, but also reflected the specific trials and experiments presented in the papers. The Keyword field also generated a detailed taxonomy, but it did not necessarily represent the focus of the articles, as can be seen with the focus on acalculia and apraxia.

APPENDIX 3 – PHRASE LINKAGES

1.High frequency phrase linked to multiple low frequency phrases

A proximity analysis was run on the themes APRAXIA and AGNOSIA for each of the four fields (Full Text, Abstracts, Keywords, and Titles). High technical content phrases in the database located in close physical proximity to the theme phrase were identified. Taxonomies of these related phrases were then constructed manually, by visual inspection. Both the number and richness of the categories for each field database are shown as follows.

Apraxia

The proximity analysis on APRAXIA on the Full Text database yielded the most categories of the Apraxia runs. The categories are: lateralization (LEFT, RIGHT); hands and limbs (HAND, LIMB, LEFT HAND, RIGHT-HANDED, RIGHT HAND); brain related (HEMISPHERE, LEFT HEMISPHERE, RIGHT HEMISPHERE, LH, LBD, RBD, LESION, LESIONS); movements (MOVEMENTS, MOVEMENT, GESTURES, GESTURE, AXIAL MOVEMENTS, AGRAPHIA); visual (VISUAL); verbal (VERBAL, LANGUAGE, APHASIA, SPEECH, APHASIC); objects (OBJECTS); motor (IDEOMOTOR, MOTOR); and sensory (TACTILE).

On the Abstract analysis of APRAXIA, the categories are: lateralization (LEFT); hands and limbs (HAND PREFERENCE, RIGHT-HANDED); brain related (HEMISPHERE, LEFT HEMISPHERE, LESIONS, LH LESIONS); movements (MOVEMENTS, MOVEMENT, AXIAL MOVEMENTS, SKILLED MOVEMENT); verbal (APHASIA, APHASIC, SPEECH); objects (OBJECTS); and sensory (AGNOSIA).

The analysis of the Keywords yielded: hands and limbs (LIMB, ARM, FINGER); brain related (CORTEX, PARIETAL, RIGHT HEMISPHERE, HEMISPHERE, MOTOR CORTEX, CORTICOBASAL DEGENERATION); movements (MOVEMENTS, LEARNED MOVEMENT, MOVEMENT, ARM MOVEMENTS); verbal (APHASIA); and motor (IDEOMOTOR).

Finally, the Title analysis yielded: hands and limbs (HANDEDNESS); and verbal (SPEECH).

Thus, not only are the numbers of phrases for each field different, but the numbers and types of categories, and the richness of detail within each category are different. Titles appeared to be the ‘weakest’ field in terms of informational content, but the inclusion of Keywords Plus in the Keywords field may provide misleading results. The message here for authors of Cortex articles is that detail in paper Titles and Keywords should not be neglected, if wide access to information is desired.

Agnosia

The proximity analysis of AGNOSIA produced similar differences among the databases.

Fulltext [only top 10% of phrases used for selection]: visual (VISUAL, SPATIAL, PERCEPTUAL, TOPOGRAPHICAL, OPTIC); objects (OBJECT, OBJECTS); auditory (AUDITORY, DEAFNESS, AUDITORY PROCESSING); association (SEMANTIC, RECOGNITION, ASSOCIATIVE, FACES, PROSOPAGNOSIA, DISASSOCIATION); memory (MEMORY, AMNESIA); lateralization (LEFT, RIGHT); deficits (IMPAIRMENT, DEFICIT, DEFICITS, DISORDER, DISORDERS, DAMAGE); verbal (VERBAL, WORD, APHASIA); tactile (TACTILE, HAND); brain (HEMISPHERE, LESION, LESIONS).

Abstract [all phrases with frequency >2 used]: visual (VISUAL, SPATIAL, IMAGERY, PERCEPTUAL, DRAWINGS, DISORIENTATION); objects (OBJECT, OBJECTS); association (RECOGNITION, SEMANTIC, FACE, FACES, NAMES, PROSOPAGNOSIA, ASSOCIATION, DISASSOCIATION); memory (MEMORY, AMNESIA); lateralization (LEFT); deficits (DEFICIT, IMPAIRMENT, DISORDER, IMPAIRED); verbal (LETTERS, ALEXIA, AGRAPHIA, ANGULAR GYRUS); tactile (TACTILE, APRAXIA); brain (LESION, SUBCORTICAL, CEREBRAL); arithmetic (CALCULATIONS, ARITHMETIC).

Keywords [all phrases with frequency >2 used]: visual (OPTIC, OPTIC APHASIA, VISUAL, IMAGERY, PERCEPTION); objects (OBJECT); association (RECOGNITION, ASSOCIATIVE, DISASSOCIATION); deficits (IMPAIRMENT); verbal (APHASIA, ALEXIA).

Titles [all phrases with frequency >2 used]: visual (VISUAL); association (ASSOCIATIVE); verbal (ALEXIA).