

## References

1. Salton, G. and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, New York, NY, 1983.
2. *FAIRS User's Manual*, GTE Laboratories, Waltham, MA, September 1987.
3. Chang, S.C. and W.C. Chen, "And-less Retrieval: Toward Perfect Ranking," *Proceedings of ASIS Annual Meeting 1987*, Oct. 1987, pp. 30-35.
4. Chang, S.C. and C. McGowan, "Full-text Retrieval in Software Maintenance," *Proceedings of COMPSAC'87*, Oct. 1987, pp. 53-57.
5. Skinner, J., "Hardware for High Speed Text Retrieval," *Proceedings of Trends and Applications*, 1984, pp. 341-347.

## STRUCTURE OF INFORMATION IN FULL-TEXT ABSTRACTS

Elizabeth DuRoss Liddy  
Syracuse University School of Information Studies  
Syracuse, New York 13244-2340

### Abstract

Although full-text documents have no explicit structure, research in discourse linguistics would suggest that since individual texts of a particular type serve a common purpose, detailed analysis of that text-type would reveal an implicit structure. An investigation was conducted into whether free-text abstracts reporting on empirical work do in fact possess a predictable, detectable structure. Following experiments with 12 expert abstractors to uncover their internalized notions as to the structure of abstracts, a detailed linguistic analysis of 276 abstracts established a text-level structure. This structure is revealed by linguistic features which can be automatically detected and used to create frame-based representations of each document. Production of structured representations of empirical abstracts is being attempted in a prototype information retrieval system currently being developed on the Connection Machine at Syracuse University.

### Introduction

I would like to respond directly to two statements in the Call for Papers indicating the 1988 RIAO Conference general theme. The statement that "the structure of the information is not known a priori" in a full-text database is not necessarily true, but perhaps not yet fully investigated. It may simply be that compared to the information stored in formatted database management systems, it is not as explicitly structured. While a full-text database may not have records with specified attributes and a limited range of values as in a DBMS, there does exist an implicit level of organization of information within each free-text record. If this structure can be adequately delineated; clues established for detecting the structure of the text processed; and a system developed for interpreting these clues and converting natural language texts into explicitly structured representations; we may well find one way to "reduce imprecision in full-text database searching".

I have been conducting research into determining whether a structure does exist in one particular text type used in full-text information retrieval, namely abstracts reporting on empirical work. As of this date, the structure has been proposed, and is currently being validated by expert abstractors. Concurrently, automatic detection and use of this structure is being implement-

d in a prototype retrieval system being developed on the Connection Machine at Syracuse University in the context of a project being conducted by R. N. Oddy and myself.

In this paper, I am interested in: 1. Communicating my rationale for posing the existence and importance of structure in this text type; 2. Suggesting how structure may be made use of in information retrieval work; 3. Presenting the results of a linguistic analysis of free-text abstracts which reveal this structure; 4. Indicating how this structure and its detection will facilitate retrieval in the system we are developing.

### Background

In current retrieval systems, document representations consist of both formatted and unformatted fields. The formatted representation consists of fields such as author, title, journal, date, and keyword descriptors. The unformatted representation is either the full-text of the document or its abstract. However, unlike the indexer who must assign keyword descriptors chosen from a pre-established list, an abstractor is free to represent a document's content with whatever words best summarize it. Although the results in two quite distinct types of document representation, the same basic techniques are used for searching both those developed for the controlled keyword representation. In current free-text systems, if a user wants documents concerned with "the effect of A on B in environment C", the retrieved set may contain documents about "the effect of C on B in environment C". Such irrelevant documents are retrieved because the representation does not include relations between concepts nor can the search mechanism require the concepts to be in the relationship needed by the user. Current free-text representations permit search techniques which merely require the desired concepts to occur or be in some particular linear order or adjacency distance within the abstract, but they cannot require the desired concepts to be in specified semantic relationships.

I am suggesting that it is possible to improve on this situation in free text retrieval by capitalizing on the special characteristics which natural language texts provide. One of the foremost things that concepts discussed in a text do not exist solitarily, but in useful semantic relations to each other which are revealed by the structure of the text. To capitalize on this fact, however, we need to learn more about the structure of the natural language texts themselves.

### Research Question

Generally stated, the hypothesis of my work is:

There is a predictable structure to abstracts reporting on empirical work which reveals the semantic roles that concepts play in the abstract, and structured document representations based on these roles can be used to improve information retrieval results.

Since they serve the same purpose, abstracts of articles which report on empirical research would be expected to exhibit much commonality in the categories of information they contain. The categories are the standard components of information which make up any research report (e.g. hypothesis, methodology, results, etc.). In addition to predictable components, the notion of a perceivable structure entails discernible connections between these components. The connections are thought to be a constricted subset of possible semantic relations or roles (e. g. causal, temporal, agent, etc.). These roles are of great importance to the task of revealing the structure of individual abstracts, because although variety exists across subject areas as to what an hypothesis, an independent variable, or the results might be, the ways in which these components relate to each other are quite restricted. And it is likely that these limited relations will be revealed in a circumscribed set of vocabulary items, referred to as lexical clues. These lexical clues (e.g. "examined influence of..."; "analyses indicate..."; "administered...to...") are hypothesized to be used consistently with minor variation, across disciplines to organize the substantive content of abstracts reporting on empirical work. In addition, it is hypothesized that recognition and interpretation of these lexical clues in abstracts is rule-governed enough to permit computer recognition and instantiation of an abstract structure representation. Since the terms or phrases which could possibly serve as lexical clues are a rather circumscribed set, a lexicon of these terms and phrases has been developed. The lexicon contains not only the semantics of the entry but also procedural information as to how surrounding text is to be fit into a structured representation of the abstract.

Furthermore, a frame-like structure (Minsky, 1975) is suggested as the most appropriate representation for organizing these components. Systems involved in text understanding tasks have found the frame structure a useful representation (Metzing, 1980). Instantiation of a frame structure for each abstract by use of lexical clues will produce a searchable structured representation which still contains the natural language of the abstract, but one which can be used to improve the level of precision obtainable using free-text searching.

My belief that there exists a discernible, predictable structure in abstracts, arises from work in discourse linguistics establishing structures of other text types, mainly narrative texts; but recently attention has turned to expository texts (Black,

1985), prompted by efforts in artificial intelligence to develop systems possessing natural language understanding capabilities. Text types have structure because conventions of form have evolved over time and it is efficient for both producers and receivers of these text types to make use of their rather fixed structure. A text-type schema aids and speeds the comprehension of texts by establishing expectations of what will follow in text, thereby allowing the receiver to concentrate on the semantic content of the individual text. Research in human text understanding has shown that knowledge of the likely structure of a text can limit greatly the otherwise enormous number of inferences which simple word and sentence processing can generate (Britton & Black, 1985).

#### Methodology

Phase I of the research (Liddy, 1987) involved 12 expert abstractors, cumulatively representing 63 years of abstracting experience, in four tasks (free-generation; typicality measure; free-sorting; and semantic relation solicitation) employing methodology similar to that used in cognitive psychology research to uncover various schemata. The results of these tasks were analyzed and used as a first approximation of the structure of empirical abstracts. This structure guided Phase II, which consisted of a detailed content analysis of a sample of 276 empirical abstracts drawn from the 1987 files of PsycINFO and ERIC. The analysis consisted of: 1) a coding of each abstract as to what components were included, and; 2) a marking of the lexical clues which served to indicate which frame slot was instantiated by each piece of text.

#### Results

Phase II analysis has produced three types of results: 1) descriptive statistics of the frequency with which each component occurred in the sample; 2) rule sets for identifying the substantive content of each component based on the expected lexical clues, and; 3) a summary of the regularity with which a list of re-occurring stems function as these lexical clues.

In Phase I the expert abstractors generated the components listed in Table 1, which are here displayed with their frequency of occurrence in the 276 abstracts analysed in Phase II. Figure 1 presents the resultant abstract structure based on the most predictable ordering of these components as observed in the abstracts. Although an average of only 9 components occur in each abstract, this fully fleshed out structure is based on the definite tendency for components to occur in particular sub-groups at certain points in the abstract. This figure may be interpreted as the base-level text-grammar of abstracts which permits transformations, similar to the more familiar sentence-level grammars. It is this proposed structure which is currently being validated by 8 expert abstractors who are performing a componential analysis of a sample of 80 abstracts.

The algorithms which will be used in our system to determine which segments of text instantiate the component slots are based on the existence of a rather circumscribed set of vocabulary items used across abstracts to present the substance of each component. As an example, Figure 2 presents a concordance of the linguistic clues which were used in the 69 instances in which the 'hypothesis' component was expressed in the abstracts. (Although Table 1 indicates that 50 abstracts contained the hypothesis component, there were multiple statements of the hypothesis in some of these abstracts.) The concordance is arranged by the stems of the lexical clues, with the rightmost column summing the multiple occurrences. The totals show that 62 out of 69 or 90% of the occurrences of the hypothesis component were detectable based on just 6 reoccurring stems.

Table 2 is a summary of similar concordances for the 29 components which had at least 10 occurrences in the sample of abstracts. These figures indicate that a total of 341 reoccurring stems can be used to detect 2800 out of 3059 or 92% of the components' occurrences.

Lexical clues will not be the only indicators as to how slots in each frame are instantiated. Texts contain other predictable characteristics which appear rule-governed enough to be captured in algorithms. These characteristics include: 1) typical order and groupings in which components tend to occur; 2) use of various verb tenses for different components, and; 3) a set of continuation clues (e.g. furthermore, in addition, also, moreover, etc) as suggested by Halliday and Hasan (1976) which indicate that the sentence in which they occur is a continuation of the component in the previous sentence. These are useful when a sentence lacks any of the component-revealing lexical clues.

#### Conclusion

Results so far indicate that the structure of free-text empirical abstracts is quite predictable and detectable. R. N. Oddy and I are currently involved in a research project at Syracuse University which will demonstrate whether, in fact, this structure can be automatically detected and used to produce structured representations of document abstracts. Such abstracts could be used in a variety of ways in a retrieval system. We will use the structured representations in an interactive retrieval system. Figure 3 shows the structured representation of one abstract which might be found in such a system, while Figure 4 provides a template of the same abstract structure plus 3 possible lexical clues for each component within each set of parentheses.

Our system, using discourse linguistic techniques similar to those applied to the abstracts, will also produce structured representations of the user's problem statement. Then, the structure of what is usually an initially poorly-defined problem statement will be used to retrieve a few seed abstracts which themselves may not solve the user's problem, but which may advance the user to a clearer view of the actual nature of his

problem. For example, the retrieved structured abstract representations may reveal to the user that the concept he is interested in has been used in other research as both an independent variable and a dependent variable, but his statement does not make it clear which role is of interest to him. Given this response, the user can interactively adjust the structured representation of the problem statement which the system presents to him. Using this restructured problem statement, the system will perform another iteration and retrieve a more precisely specified set of abstracts. The process could continue over several iterations, with the user adjusting his statement's representation in reaction to the structured abstracts he is shown until he has eventually retrieved a set of abstracts in which the concepts of interest do exist in the roles appropriate to his need.

Although this paper reports on the investigation of structure in just one text-type, the results suggest that given the approaches and methodologies which the burgeoning field of discourse linguistics has to offer, the possibility exists that full-text documents do not necessarily need to remain outside the domain of those documents which can be handled successfully in information retrieval systems.

Table 1: Components Observed in Sample of Abstracts

COMPONENT	PsycINFO		ERIC		TOTAL	
	#	%	#	%	#	%
subjects	125	.99	136	.91	261	.95
findings/results	115	.91	133	.89	248	.90
purpose	99	.79	139	.93	238	.86
research topic	76	.60	123	.82	199	.72
references	126	1.00	36	.24	162	.59
procedures	90	.71	68	.45	158	.57
data collection	33	.26	95	.63	128	.46
conclusions	66	.52	55	.37	121	.44
independent variable	35	.29	41	.27	76	.28
dependent variable	34	.27	41	.27	75	.27
methodology/research design	25	.20	49	.33	74	.27
conditions/treatments	50	.40	16	.11	66	.24
location of study	1	.008	64	.43	65	.24
relation to other research	42	.33	18	.12	60	.22
data analysis	15	.12	40	.27	55	.20
hypothesis	35	.28	15	.10	50	.18
implications for practice	14	.11	33	.22	47	.17
number of experiments	42	.33	3	.02	45	.16
background	15	.12	28	.19	43	.16
research question	15	.12	26	.17	41	.15
time frame of study	6	.05	32	.21	38	.14
tables included	0	.00	36	.24	36	.13
appendices included	0	.00	35	.23	35	.13
sample selection tech.	5	.04	23	.15	28	.10
discussion	14	.11	9	.06	23	.08
intended use/prac. applic.	2	.02	12	.08	14	.05
control population	5	.03	7	.05	12	.04
significance of findings	4	.03	8	.05	12	.04
future research needs	1	.008	9	.06	10	.04
new terms defined	5	.04	4	.03	9	.03
materials	2	.02	4	.03	6	.02
limitations	0	.00	3	.02	3	.01
unique features of study	3	.02	1	.006	4	.01
administrators of study	0	.00	2	.01	2	.007
institution doing study	0	.00	2	.01	2	.007
reliability of findings	0	.00	2	.01	2	.007
tests	0	.00	1	.007	1	.004
apparatus	0	.00	0	.00	0	.00
confounding variables	0	.00	0	.00	0	.00
drugs administered	0	.00	0	.00	0	.00
intended audience	0	.00	0	.00	0	.00
scope	0	.00	0	.00	0	.00
subsequent research planned	0	.00	0	.00	0	.00

Figure 1: Structure of Empirical Abstracts

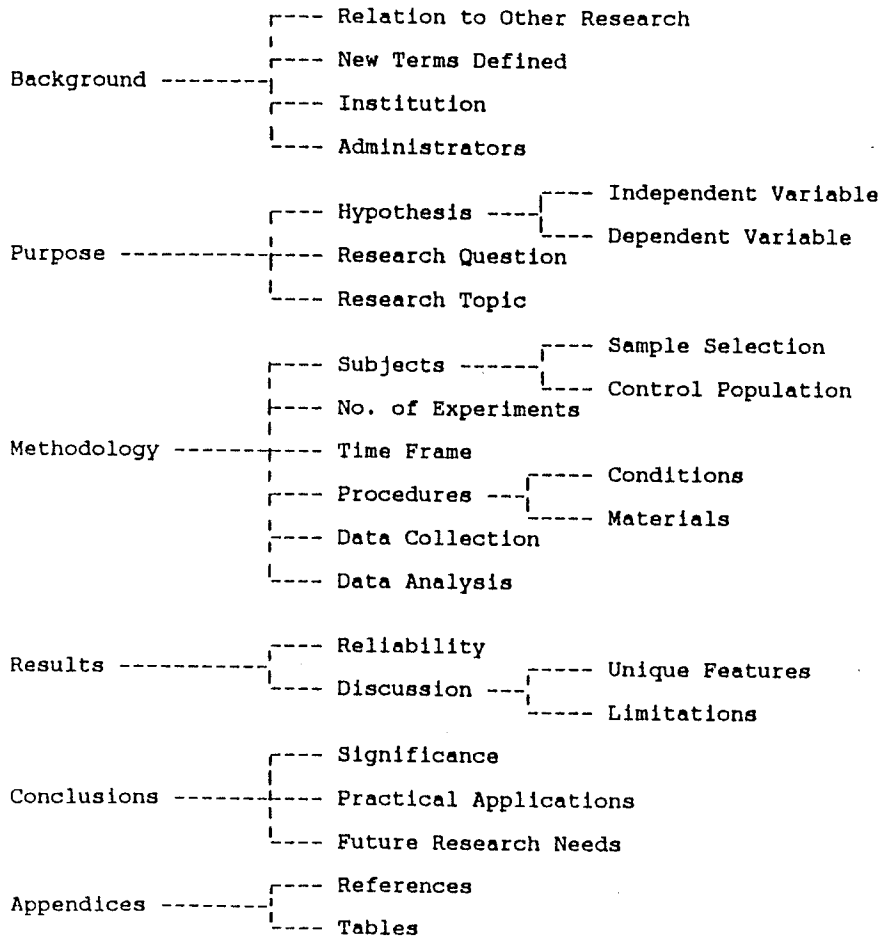


Figure 2: Concordance of 'Hypothesis' Clues

ARGUED that _____	2	2
it was ASSUMED that _____	1	
it was CONJECTURED that _____	1	
original EXPECTATION that _____	1	2
as EXPECTED, _____	1	
HOLDS that _____	1	
[N] HYPOTHESES being ... that _____	1	32
[N] HYPOTHESES were advanced: _____	1	
to test / [N] / HYPOTHESES: _____	2	
confirm all HYPOTHESES in that _____	1	
supportive of the HYPOTHESES, _____	1	
supporting the HYPOTHESIS of _____	1	
the HYPOTHESIS that _____	14	
as HYPOTHESIZED, _____	1	
HYPOTHESIZED that _____	10	
IF _____ THEN _____	1	
POSSIBILITY that _____	1	
POSTULATES that _____	1	
PREDICTED _____	1	21
a PREDICTED _____	1	
PREDICTED, ..., that _____	1	
as PREDICTED _____	7	
it was PREDICTED that _____	3	
_____. To test this PREDICTION _____	1	
PREDICTION that _____	4	
in accordance with PREDICTIONS _____	1	
PREDICTS that _____	2	
PROPOSING that if _____	1	2
PROPOSITION that _____	1	
_____. This notion is TESTED.	1	3
TESTED the claim that _____	1	
TESTED the theory that _____	1	
consistent with the VIEW that _____	1	
TOTALS	69	62

Table 2: Rank by Percentage Accounted for by Multiple Stems

COMPONENT	Number of Stems Occurring More than Once	Number of Occurrences Accounted for by these Stems	Percentage Accounted for by these Stems
References	3	156/156	100 %
Appendices	3	30/30	100 %
Number of Experiments	2	44/45	98 %
Research Question	5	42/43	98 %
Methodology/Research Design	13	78/81	96 %
Subjects	28	297/311	95 %
Location	7	61/64	95 %
Research Topic	20	188/199	94 %
Data Collection	17	156/166	94 %
Data Analysis	8	59/63	94 %
Tables	1	34/36	94 %
Findings/Results	73	544/589	92 %
Purpose	20	265/287	92 %
Control Population	2	11/12	92 %
Relation to Other Research	10	63/69	91 %
Conditions/Treatments	13	93/103	90 %
Hypothesis	6	62/69	90 %
Future Research Needs	2	9/10	90 %
Time Frame	4	32/36	89 %
Conclusions	16	145/164	88 %
Implications	6	47/54	87 %
I. V. / D. V.	13	78/91	86 %
Sample Selection Technique	4	24/28	86 %
Procedures	48	209/250	84 %
Practical Applications/Use	4	15/18	83 %
Significance of Findings	4	9/12	75 %
Background	8	34/49	69 %
Discussion	1	15/24	62 %
<b>TOTALS</b>	<b>341</b>	<b>2800/3059</b>	<b>92 %</b>

Figure 3: Structured Representation of an Empirical Abstract

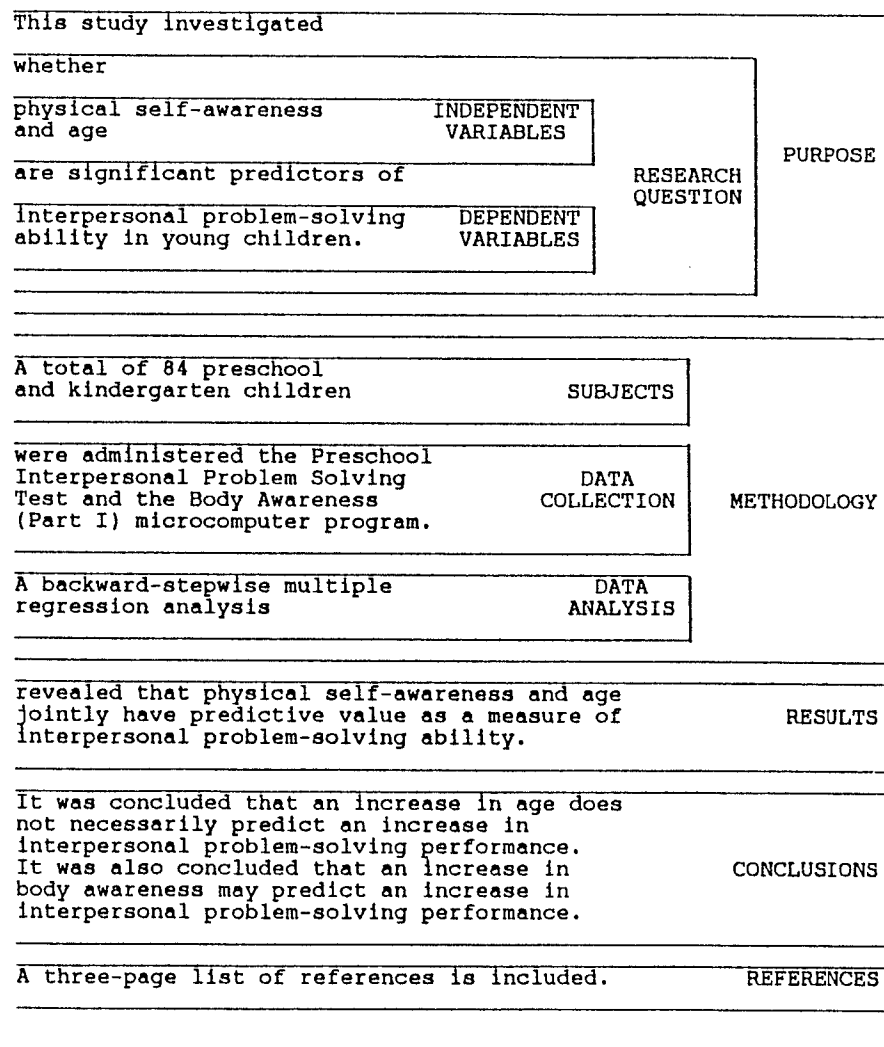


Figure 4: Generalized Template of Structure with Sample Clues

[investigated / examined / assessed]		PURPOSE
[whether / if / how]		
INDEPENDENT VARIABLES	RESEARCH QUESTION	
[are significant predictors of / effect on / influences]	DEPENDENT VARIABLES	
[A total of ___ / groups of ___ / samples of ___ ]		METHODOLOGY
SUBJECTS		
[were administered ___ / completed a questionnaire on ___ / [ ___ were assessed ]]	DATA COLLECTION	
[analyzed by ___ / were compared using ___ / ___ was calculated]	DATA ANALYSIS	RESULTS
[revealed that ___ / demonstrated that ___ / ___ affected ___ ]		
[It was concluded that ___ / findings support ___ / data suggest ___ ]		
[Appendices include ___ / ___ are appended / tables of ___ ]		APPENDICES

References

- Black, J. (1985). An exposition on understanding expository text. In B. Britton & J. Black (Eds.), Understanding expository text: A theoretical and practical handbook for analyzing explanatory text (pp.249-67). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Britton, B., & Black, J. (1985). Understanding expository text: From structure to process and world knowledge. In B. Britton & J. Black (Eds.), Understanding expository texts: A theoretical and practical handbook for analyzing explanatory text (pp.1-9). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. London: Longman Group Ltd.
- Liddy, E. D. (1987). Discourse-level structure in abstracts. In Proceedings of the 50th Annual Meeting of the American Society for Information Science. (pp. 138-47). Medford, NJ: Learned Information.
- Metzing, D. (Ed.). (1980). Frame conceptions and text understanding. New York: Walter de Gruyter.
- Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), The psychology of computer vision (pp. 11-77). New York: McGraw-Hill.