DOCUMENT RESUME

ED 307 884 IR 052 772

AUTHOR Liddy, Elizabeth D.
TITLE Discourse-Level Structure in Abstracts.
PUB DATE Oct 87
NOTE 11p.; Paper presented at the Annual Meeting of the American Society for Information Science (Boston, MA, October 4-8, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Abstracts; *Componential Analysis; *Discourse Analysis; Information Retrieval; Matrices; *Syntax; *Text Structure; Users (Information)
IDENTIFIERS *ERIC; *PyschInfo

ABSTRACT

An investigation was undertaken into the possibility of automatically detecting how concepts exist in relation to each other in abstracts, a text-type commonly used in free-text retrieval. The end goal of this research is to capture these relationships in structured representations of abstracts' contents so that users can require not only that the concepts of interest to them co-occur in the retrieved documents, but also that the roles they play in relation to one another are the ones of interest. Four tasks found useful in revealing other schema were performed by expert abstractors. The results were analyzed and used as the basis for developing a frame-like structure of abstracts reporting on empirical work. A discourse linguistic analysis of a sample of 276 abstracts identified the lexical/syntactic clues which could be used by a system to automatically instantiate the frame-like structure of individual abstracts. The text is supplemented by four tables and three figures. (10 references) (Author)

DISCOURSE-LEVEL STRUCTURE IN ABSTRACTS

Elizabeth D. Liddy

Syracuse University, School of Information Studies, Syracuse,

Abstract. An investigation was undertaken into the possibility of automatically detecting how concepts exist in relationship to each other in abstracts, a text-type commonly used in free-text retrieval. The end goal of this research is to capture these relationships in structured representations of abstracts' contents so that users can require not only that the concepts of interest to them co-occur in the retrieved documents, but also that the roles they play in relation to each other are the ones of interest. Four tasks found useful in revealing other schema were performed by expert abstractors. The results were analyzed and used as the basis of developing a frame-like structure of abstracts reporting on empirical work. A discourse linguistic analysis of a sample of 276 abstracts identified the lexical/syntactic clues which could be used by a system to automatically instantiate the frame-like structure of individual abstracts.

OVERVIEW

While free-text searching has improved to some extent an information system's ability to retrieve only those documents of interest to a user, it still does not produce results sufficiently refined for those users who can specify quite precisely what the content of relevant documents should consist of. This is because current free-text retrieval permits users to require only that concepts of interest to them co-occur in a document. As a result, many nonrelevant documents are retrieved, because the search mechanism cannot require the concepts to be in the relationship needed by the user [1]. And although there are search techniques which require the desired concepts to be in some particular linear order or adjacency distance within the abstract, there are none that require the desired concepts to be in specified semantic relationships.

In an attempt to improve on this situation, an investigation was undertaken into the possibility of automatically detecting how concepts exist in relationship to each other in empirical abstracts, a text-type commonly used in free-text retrieval. The goal of this research is to capture these relationships in structured representations of abstracts' contents so that users can request not only that concepts of interest occur in the retrieved documents, but also that these concepts exist in the desired semantic relationships.

BACKGROUND

The belief that a structure exists in abstracts arises from work done in discourse linguistics, which is concerned with the study of units of language larger than a sentence. These larger units are referred to as texts, and have been the focus of increasing study in linguistics, artificial intelligence and natural language processing. One line of investigation in discourse linguistics has been the detection of a particularized structure within a given text type. Text types found to exhibit characteristic syntactic and semantic organization with predictable consistency within that type include folk tales [2], narratives [3], and scholarly papers [4]. The research being reported here has extended this line of investigation by discovering and delineating the structure of the text-type of empirical abstracts.

The theoretical basis of this work derives partially from research done in cognitive science showing that human understanding requires efficient schemas for the organization of knowledge. One of the most widely accepted knowledge organizing theories is Minsky's frame structure theory [5]. A frame is a learned data-structure originally proposed as a formalism for explaining human vision and later used for describing human memory. The frame formalism has been useful in research in human text understanding and has been successfully extended for use in a variety of computerized text understanding systems (see [6] for examples).

The current study suggests that in the same way that a frame serves as a formalism for representing text type structures in memory, a frame structure can be detected in the text itself. In addition, the investigation was concerned with showing that the specific lexical clues which indicate to humans how to instantiate their mental frame of a particular text type are rule-governed enough to permit automatic instantiation of a frame structure for individual empirical abstracts.

A structure consists of components and the relations among them. In text structure, the components are those necessary categories of text content which define the text type. Relations are properties that hold between two or more entities and define the type of interaction, influence or simply co-occurrence that holds between the entities.

METHODOLOGY

The question of whether there is a predictable, framelike structure in abstracts reporting on empirical work, was investigated by tapping the expertise of professional abstractors to delineate the components and relations which comprise the abstract frame structure. This was done by means of four tasks employing methodology similar to that used in cognitive psychology research to uncover various schemata [7, 8, 9].

Task 1, a free-generation task, was administered by mail to 14 professional abstractors from either ERIC or PsycINFO. These abstractors were simply asked to list all the components of information that are included in an abstract of an empirical study. For the remaining tasks, each subject used the complete list of components generated by all the abstractors from their respective service.

Tasks 2, 3 and 4 were administered in person at the facility of each abstractor. The tasks were administered in small groups of two to four subjects and the three tasks took a total of about 1 and 1/2 to 2 hours of a subject's time.

Task 2 asked the subjects to first indicate which of the components in the list were, to their way of thinking, the most typical of an empirical abstract. They were to then go back through the list and mark the components they considered to be of the next level of prototypicality. This process was to be continued as long as the subjects felt there were differences in degree of typicality.

In Task 3, each subject was given a pack of cards, each card containing the name of a component from the list used in Task 2, plus written instructions for a multiple sorting procedure. A multiple sorting procedure simply asks subjects to assign elements to categories of their own choosing [10]. The value of the procedure is that no preconceived limitations are set on how the subject is to perform the sort. The method is ideal for this research, since it allows the subject to impose whatever structure they desire on the components.

Subjects were asked to spread the cards out and then sort them into groups in such a way that all the cards in each group had something in common. Subjects were allowed to perform as many different sorts as they wanted.

Finally, Task 4 served to identify the semantic relations comprising the frame structure of empirical abstracts. Subjects were instructed to draw lines from one component to the other components with which, in their opinion, there was a relationship and to write on the connecting line some word or words to describe that relationship.

RESULTS

The components freely generated in Task 1 were normalized so that synonymous ways of referring to the same component were reduced to a canonical term or phrase. Abstractors from PsycINFO generated 24 components and the abstractors from ERIC generated 35 components, with 15 of these components common to both groups of abstractors. Table 1 contains all the components generated with the number of abstractors who suggested each component.

Of the ten ERIC abstractors who participated in Task 1, only eight were available to participate in Tasks 2-4, while all four abstractors from PsycINFO participated. The results from these abstractors on Task 2 produce the ranked ordering of components of an empirical abstract and their typicality scores seen in Table 2. The subjects' original typicality values were reverse coded and then converted to proportions so that all components judged as being at the highest level of typicality equal 1 no matter how many levels of typicality an individual judge may have used. These scores were then averaged and the averages for the 15 components mentioned by both sets of abstractors were summed.

As can be seen from comparing the ordering of the 15 common components based on typicality ratings in Table 2 with the ordering based on frequency of free generation of components in Table 1, having subjects assign typicality scores to a prepared list of components changes the relative ordering to some extent. This is not surprising, however, since recall and recognition are known to be very different memory tasks and a component which was simply not recalled by an individual abstractor in the free generation task may later be recognized as quite typical of an empirical abstract.

Table 3 presents a final ranked ordering of the 15 common components based on the combined results of Task 1, the free-generation task, and Task 2, the typicality rating task. Although these tasks are admittedly different in nature, the rankings in Table 3 present

a preliminary indication of the relative significance of these components in the mental framework of this group of expert abstractors.

From Task 3, the free-sorting task, only the results based on one type of sort, the grouped-ordering sort are reported here. This was the most commonly used scheme for sorting (10 out of 12 subjects) and also a source of essential information in constructing a predictable frame structure. Sorting on this parameter provided not only the higher level structuring of empirical abstracts but also information as to which components co-occur within each of these 'meta-components'.

For illustration, the sort of one subject, who made and orally labeled five piles of cards is presented in Figure 1. Listed beneath each pile's label are the abstract components designated by the subject as belonging to that group.

Using the grouped-ordering sorts of the 10 abstractors, matrices of the frequency with which each of the 15 common components was placed in the same group as every other component were constructed for 1) ERIC, 2) PsycINFO and 3) a composite of both. The composite matrix is presented in Table 4.

Figure 2 is a graphic representation of the 15 common components using the matrix values in Table 4. This representation, which is to be read clockwise from the upper left-hand corner, is intended to convey more clearly a notion of the basic structure existing within such abstracts. The lines encircling the three groupings are arbitrarily sketched, but can be seen to enclose sets of components which exist in very strong and inter-connected associations with each other.

The results of Task 4, which asked abstractors to specify the relations they see as existing among abstract components, were quite extensive and will not be presented here in their entirety. Figure 3 does serve to suggest the type of relations offered by abstractors by adding to each link a lexical expression of one semantic relation offered by abstractors.

CONCLUSIONS

The nature of an abstract's frame structure uncovered in the results of the four tasks reported above is currently being used to guide the search for rules governing the ways this structure is revealed by lexical clues. In order to demonstrate that the frame structure of empirical abstracts can be useful in information retrieval tasks, it is essential to show that this structure can be automatically detected, and a frame structure actually instantiated for each individual empirical abstract processed. Ongoing research will show how the guidance offered by the expert-generated structure was used to develop lexical clue recognition rules and how these rules, when applied to a sample set of empirical abstracts, produce structured representations.

Results of the next stage of the research which is currently nearing completion will indicate whether rule-governed instantiation of the abstract frame structure can be accomplished. Positive results would support the feasibility of automatic processing of abstracts to fill the slots of an abstract frame. Automatic instantiation would produce a representation containing not only the substantive content of an abstract's components but also indicating which frame component the information belongs to and how this information is related to other information in the abstract. Such representations offer the potential for producing retrieval results of greater precision.

NOTES

1. C. Borgman, D. Moghdam & P. Corbett, Effective Online Searching (New York: Marcel Dekker, 1984).

2. V. Propp, Morphology of the Folk-tale (L. Scott, Trans.). (Bloomington: Indiana University Press, 1958). (Original work published 1929).

3. R. Longacre, The Grammar of Discourse (New York: Plenum Press, 1983).

4. T. A. van Dijk, Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition (Hillsdale, NJ: Lawrence Erlbaum Associates, 1980).

5. M. Minsky, "A Framework for Representing Knowledge." In P. Winston (Ed.), The Psychology of Computer Vision (New York: McGraw-Hill, 1975), 11-77.

6. D. Metzing (Ed.), Frame Conceptions and Text Understanding (New York: Walter de Gruyter, 1980).

7. G. Bower, J. Black & T. Turner, "Scripts in Memory for Text," Cognitive Psychology, 11 (1979), 177-220.

8. N. Cantor, "A Cognitive-Social Approach to Personality." In N. Cantor & J. Kihlstrom (Eds.), Personality, Cognition, and Social Interaction (Hillsdale, NJ: Lawrence Erlbaum Associates, 1981), 23-44.

9. A. Graesser & S. Goodman, "Implicit

knowledge, Question Answering, and the Representation of Expository Text." In B. Britton & J. Black, (Eds.), Understanding Expository Texts: A Theoretical and Practical Handbook for Analyzing Explanatory Text (Hillsdale, NJ: Lawrence Erlbaum Associates, 1985), 109-171.

10. D. Canter, J. Brown & L. Groat, "A Multiple Sorting Procedure for Studying Conceptual Systems." In M. Brenner, J. Brown & D. Canter (Eds.), The Research Interview: Uses and Approaches (London: Academic Press, 1985), 79-114.

Table 1: Frequency of Component Generation

COMPONENT	ERIC (N=10)	PsycINFO (N=4)	Total (N=14)
GENERATED BY BOTH SERVICES			
hypothesis	10	3	13
subjects	9	4	13
methodology	8	3	11
findings	7	3	10
results	8	2	10
purpose	4	4	8
conclusions	4	3	7
relation to other research	4	3	7
implications	5	2	7
discussion	3	2	5
references	2	2	4
conditions/treatments	1	2	3
sample selection technique	1	2	3
intended use/practical applications	2	1	3
research design	1	1	2
ERIC ONLY			
future research needs	7		7
data analysis	4		4
institution doing study	4		4
location of study	4		4
time frame of study	4		4
appendices included	3		3
dependent variable	3		3
independent variable	3		3
administrators of study	2		2
background	2		2
confounding variables	2		2
intended audience	2		2
tables included	2		2
data collection	1		1
limitations	1		1
new terms defined	1		1
reliability of findings	1		1
subsequent research planned	1		1
unique features of study	1		1
PsycINFO ONLY			
tests		4	4
drugs administered		3	3
procedures		3	3
apparatus		2	2
significance of findings		2	2
control population		1	1
materials		1	1
number of experiments		1	1
research question		1	1
scope		1	1

Table 2: Rankings Based on Averaged Typicality Scores

COMPONENT	ERIC	PsycINFO	TOTAL
COMMON TO BOTH SERVICES			
methodology	1	1	2
findings	.975	1	1.975
results	.950	1	1.950
purpose	.944	1	1.944
hypothesis	.938	1	1.938
subjects	.925	1	1.925
conclusions	.975	.938	1.913
research design	.901	.938	1.839
references	.576	1	1.576
sample selection technique	.598	.915	1.513
discussion	.791	.56	1.351
intended use/practical applications	.739	.56	1.299
implications	.72	.56	1.28
relation to other research	.589	.642	1.231
conditions/treatments	.498	.688	1.186
ERIC ONLY			
data collection	.851		.851
unique features of study	.788		.788
data analysis	.77		.77
time frame of study	.765		.765
background	.76		.76
dependent variable	.749		.749
tables included	.701		.701
independent variable	.696		.696
appendices included	.67		.67
intended audience	.639		.639
future research needs	.625		.625
institution doing study	.622		.622
limitations	.599		.599
location of study	.592		.592
confounding variables	.549		.549
reliability of findings	.499		.499
subsequent research planned	.49		.49
administrators of study	.485		.485
new terms defined	.448		.448
PsycINFO ONLY			
control population		1	1
drugs administered		1	1
number of experiments		1	1
research question		1	1
tests		1	1
procedures		.915	.915
significance of findings		.83	.83
apparatus		.705	.705
scope		.645	.645
materials		.498	.498

Table 3: Ranking Based on Tasks 1 & 2

COMPONENT	TASK 1 RANK	TASK 2 RANK	SUM OF RANKS	FINAL RANK
methodology	3	1	4	1
findings	4.5	2	6.5	2.5
hypothesis	1.5	5	6.5	2.5
results	4.5	3	7.5	4.5
subjects	1.5	6	7.5	4.5
purpose	6	4	10	6
conclusions	8	7	15	7
references	11	9	20	8
discussion	10	11	21	9.5
implications	8	13	21	9.5
relation to other research	8	14	22	11
research design	15	8	23	12.5
sample selection technique	13	10	23	12.5
intended use/practical applications	13	12	25	14
conditions/treatments	13	15	28	15

Subject 4 - PsycINFO

RESEARCH QUESTION	SUBJECT POPULATION	METHODOLOGY
research question	no. of experiments	methodology
hypothesis	sample selection	apparatus
scope	subjects	procedures
purpose	control population	materials
		research design
		conditions
		tests
		drugs administered

FINDINGS	RESULTS APPLIED
results	practical applications
findings	implications
significance	relation to research
conclusions	
discussion	

Figure 1: Example of One Grouped-Ordering Sort

Table 4: Co-occurence of Components in Same Group

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. methodology															
2. findings															
3. hypothesis															
4. results		9													
5. subjects	7		2												
6. purpose			8												
7. conclusions		8		9											
8. references															
9. discusion		7		6			7								
10. implications		3		2			3		6						
11. relation to research			2			4	2		2	4					
12. research design	9		2		6										
13. sample selection	7				9							6			
14. intended use						3	2		3	6	6				
15. conditions	5				3							5	4		

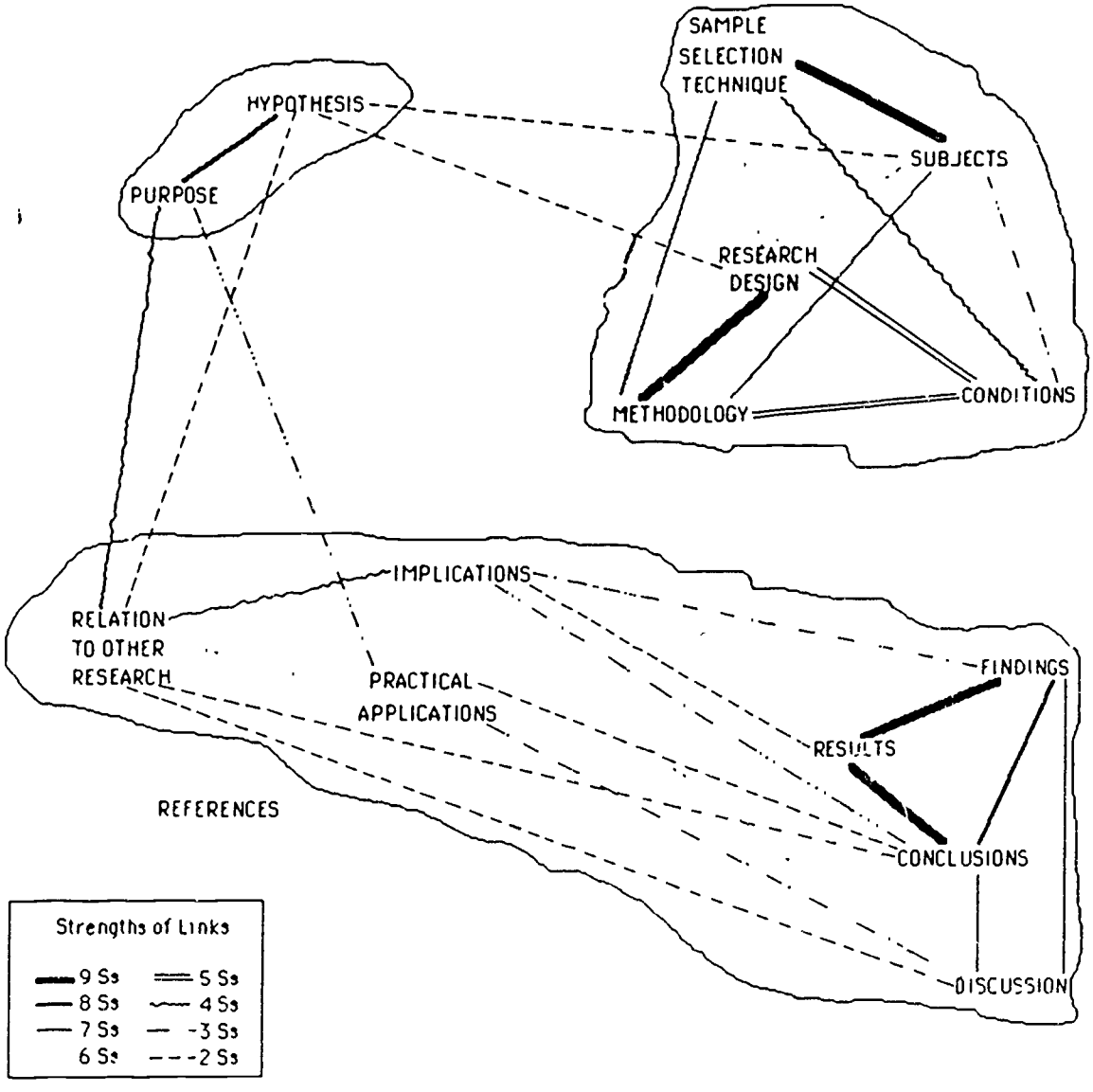

Figure 2 Strengths of Relations Between Components

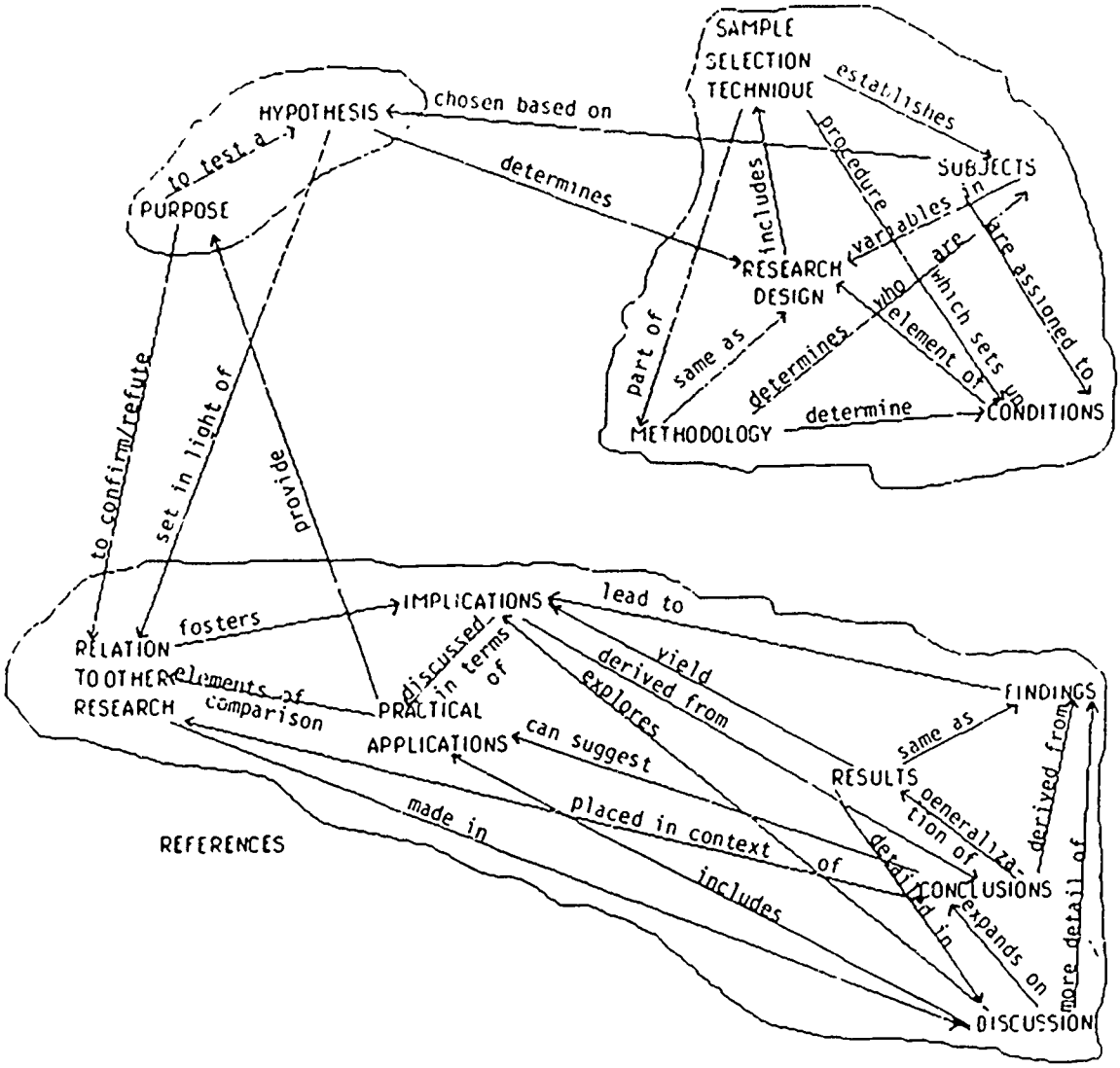

Figure 3 Sample of Relations Between Components