A META-INFORMATION STRUCTURE

FOR REPRESENTING ARGUMENTS IN SCIENCE TEXT

By

Philip John Sallis

A thesis submitted for the degree of Doctor of Philosophy
to The City University London, Centre for Information Science,
where the research was conducted

November 1979

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

ACKNOWLEDGEMENTS

There are a number of individuals to whom I would like to formally express my thanks for their support during the research and writing of this thesis.

For allowing me to carry out the research at the Centre for Information Science, I would like to thank the Director, Dr. R.T. Bottle. Also to those members of his staff who have taken an interest in my work; in particular Dr. Stephen Robertson and Verina Horsnell who has since left the Centre. She supervised my Masters year and advised me during the preparation of the initial proposal for this research; she has remained an interested observer and close friend. I owe a special thanks to my supervisor Dr. Nicholas Belkin. As well as becoming a good friend, he has been a strong influence on my work, teaching me the value of discussing problems from first principles whilst arguing with me over what those principles are. His work has made a substantial contribution to my thinking and consequently this thesis.

I would like to thank two former colleagues who have given me day-to-day support and encouragement; John Eyre and Ellen Gredely. At times both have willingly given assistance in matters connected with this research. Also to B.C. Brookes who has given me constructive criticism and encouragement during crucial phases of the development of my theory.

I am grateful to Mr. Peter Cope and the Editor of the Journal of the British Astronomical Society for permission to reprint his paper, 'The Moon Illusion', which I used as the text in two of my experiments. And to the Editor of the Journal of Informatics for permission to reprint Mr. E. Michael Keen's paper, 'The need for psycholinguistic testing to solve practical problems in information retrieval', which I also used.

A big debt of thanks goes to Kay Nicholas for typing this thesis in such an excellent manner and to Dave Nicholas for his comments and proof-reading.

Finally, thanks to all my family and friends who have endured my obsession with this research over the past three years; in particular my wife Patricia and son Adrian who are now allowed entry to the room at the top of the stairs.

DECLARATION OF COPYRIGHT

I hereby declare that until any time in the future when this declaration may be revoked, the Librarian at The City University London may use discretion for this thesis to be copied either in full or in part for the purposes of further research or academic pursuit without reference to me. If any part of this thesis is copied as aforesaid, due acknowledgement to the author should be made.

ABSTRACT

The research for this thesis has been concerned with defining and demonstrating the existence of certain semantic elements in English natural language science text which can be called meta-information. Meta-information is described as being the organisational, rather than the conceptual properties of an author's 'message' in text. Conceptual information is that subject-related output from a document which readers assimilate or synthesise with their current state-of-knowledge. Meta-information reflects the organisation or structural format used by an author to present conceptual information for transfer from text to readers. The example used here to demonstrate the existence of meta-information, is a format for the presentation of empirical argument in science text. At its most simple, a meta-informational element could be a report section-heading like, INTRODUCTION, which describes (we assume), the contents of the subsequent text. At a lower level of analysis the phrase, 'This paper describes ...', contains some semantic inference that the complete statement is one of an introductory nature; therefore, such a statement could be labelled as one of INTRODUCTION for meta-informational purposes. A 'grammar' or set of meta-informational elements, has been developed as a means of identifying certain semantic aspects of text. This grammar is based on some experimental evidence and the consensus view of readers and writers of science text who produced what has been called a conventional format for empirical argument presentation. An initial set of rules for implementing this grammar have also been developed. The rules have been tested for replicability with positive results. Although analysis of full text has shown deviation from a 'conventional argument structure', readers' summaries of the same text conform to this structure. Thus, a model of the phenomenon of information transfer from text to readers, which includes a structural transformation process based on the experimental results, has been built. A computer simulation is given to demonstrate the model in an inter-active program-user system designed to produce

summaries of whole text. The thesis is that evidence exists for the presence of meta-information in science text and that if a grammar appropriate to the kind of output information required by users is built, highly structured text could be produced so that the process of information transfer is optimised.

CHAPTER I

INTRODUCTION

1.1 OVERVIEW OF THIS CHAPTER

The intention of this chapter is two-fold. First, it is necessary right at the outset of the thesis to give an introduction to the topic being discussed and to show its relevance to the field of Information Science. This endeavour I hope has been successful. Second, I have outlined the principal assumption on which the research is based. There is no hypothesis as such, although one may be inferred from the evidence which is given to suggest the assumption. The broad aims of the research have been described together with a statement of purpose. Finally, rather than describe in detail the research methodology, I have outlined a brief summary of work done and pointed to methods and procedures in context where they arose.

1.2 RESEARCH PROBLEMS FOR INFORMATION SCIENCE

Problems in Information Science seem to be characterised by being two-faceted entities. On the one hand we deal with physical quantities such as numbers of documents and data from documents, like the distribution of citations in a particular subject journal. On the other hand we face problems such as defining readers' information needs and of meaning, understanding and problems of an abstract or conceptual nature. Many of our problems seem to combine these two broad characteristics.

If we regard information itself as being fundamental to the communication and thus furtherance of knowledge in all subjects, we could propose that to solve problems associated with its nature -

and organisation requires a corpus of knowledge and methodology all of its own. In my view, that is the first real problem of the 'science' of information. As yet there is no integrated or conventional methodology for the study of problems in Information Science. Some problems lend themselves to the methodology of say the social sciences where data can be collected and analysed in some comparative sense to produce sets of quantified results from which inferences or trends can be seen. The distribution of journal citations across a particular subject can indicate trends of research within that subject, for instance. Other problems in Information Science lend themselves more to a humanistic approach for their investigation, or the methodology of Philosophy - although I disagree with a recent claim that Information Science should be considered an extension of Philosophy (see MARTYN 1978). Problems which do come within tnis latter methodology are those which deal with the fundamental nature of the discipline itself and try to define terms and concepts for the study of Information Science.for instance. The point I am trying to make is that as an emerging discipline and one which is fundamental to the communication of all knowledge, Information Science should be and is developing methods of studying problems of information outside of any particular subject field. In this thesis however, I will be concentrating on a problem associated with science text, rather than documents relating to the humanities or literature. Also, I am primarily concerned with the English Language and all of my theoretical assumptions relate to it. This research endeavours to contribute to a methodology which states a theory based on assumptions and previously established concepts, then demonstrates the theory within the context of the universe to which it refers. In short, this is an axiomatic methodology. Some experimentation has taken place and will be described later in the thesis, but this has been carried out in order to observe phenomena arising from previously held philosophical notions, rather than for empirical evidence gathering. Writing about the transformation of research findings into scientific knowledge, GILBERT (1976) expresses the belief that when enquiring into the everyday practices of research scientists, we should cast ourselves into an interpretative rather than normative paradigm. This is how I believe we should investigate those problems in Information Science which often involve qualitative rather than quantitative evaluations of human behaviour. This thesis represents just this sort of research. I am concerned with the phenomenon of human readers interpreting

information which suggests to them the organisation of authors' arguments in text. Any sample of humans I take to indicate how this process of information transfer works will always be insufficient and inadequate to represent what everybody does all of the time. Therefore, I have tried to get as near to the 'truth' as possible and to interpret what those individuals have done. If certain commonalities arise from this observational evidence, then some theory of what might take place in a general sense can be developed. GILBERT (1976) says one more thing which is relevant to my problem-solving approach for this research. In his paper he was concerned with showing how research findings are transformed into accredited factual knowledge and he ends up by demonstrating that an important aspect of persuasion in convincing others of a theory, is the use of a familiar structure for presenting results. That is, if one can relate a new theory or concept to those already held by the listener, the validity of the new results will become more apparent more quickly. It is for this reason that I have chosen the organisation of authors' arguments in science text as the example from which to develop my theory of information transfer. The way in which individuals present empirical arguments in text is so generally accepted that I have called it the conventional format for arguments. My 'proof' for this assumption is discussed later in the thesis. Not only does familiarity with concepts such as this help others to relate to the eventual theory, but it also makes investigation of the problem easier because individuals can give immediate answers to questions about the process of reading and interpreting text if they are familiar with the concept being discussed.

It seems that there are two fundamental questions which require discussion at the outset, before a problem under review can be seen as related to the study of Information Science. First, where do problems which appear to be inherently related to the 'science' of information exist, in relation to research within other disciplines - particularly Computer Science, Psychology, Linguistics and Artificial Intelligence; all of which are concerned to some extent with the communication of information as a topic for investigation. BROOKES (1978) has suggested that Information Science may become the 'foundation science' for all the social sciences; much the same as Physics is for the natural sciences. If this is so then we are in urgent need of finding some common denominators for problem-solving and should be establishing a

greater number of definitions from which to develop working research methodologies. If our interpretation of terms such as information itself are going to be accepted by other disciplines, then we must show their general applicability to problem solving and solution descriptions. This discussion inevitably brings us to the debate on whether or not Information Science is a 'true science'. I have already mentioned one opinion that Information Science should be considered an extension of Philosophy and one which is more pragmatic that states that the discipline should be thought of as fundamental to all the social sciences. YOVITS (1969) produced an extensive survey of various concepts of science and gave a comprehensive rationale for Information Science being recognised as a true scientific discipline. He considers that the nature of problems in Information Science are so different from those of say History, that we need analytical methods and approaches to problem-solving which are of a more integrated and symbolic type. This assumption suggests techniques that are more like those of the physical sciences. This topic has been the subject of a great deal of discussion over recent years, (see particularly Anthony DEBONS (1974), Information Science: search for identity), and I will not enter further into the debate here. In my view, the individuality or characteristic strength which Information Science will attain, must come from our actually creating our own concept definitions and establishing well-defined and integrated methodologies and demonstrating their use. Information Science has I feel, the task of demonstrating theories and concepts which go beyond the bounds of one particular discipline and must relate fundamentally to the nature and being of all. My topic has inherent properties of linguistics, philosophy, psychology, and computer science, but essentially relates to Information Science, because the emphasis is placed upon establishing a theory regarding the organisation of information in text.

The second question which we should ask ourselves about problems which are considered as being potentially relevant to the study of Information Science, is just what are our present assumptions, premises and definitions concerning the nature and problems of information? To some extent this question must be examined selectively. That is to say that we only choose those assumptions, definitions and so on which relate to the problem in hand. In this research I am most concerned with the organisation of information in science text and how that information is transferred to a human reader. I am being selective here in that

I am not considering text in the humanities or literature and I am concerned with a writer-text-reader communication system, rather than merely on the level of human-to-human for instance. For my topic I have examined the nature of information itself, human communication systems and the phenomenon of information transfer from text to readers.

The fundamental problems of Information Science then, centre around explanations of the nature of information and its organisation. A particular aspect of this general problem area is the description of different information structures which are assumed to be associated with text. A recent piece of research in Information Science has shown how information in text can be directly related to the conceptual structure of the originator of that text. This is the work of BELKIN (1977) which is described later in this thesis. I will show how his work has helped me to form a terminological and conceptual base for the study of my problem.

It is assumed in this thesis that the organisation of information in text is related to a system that is intended to communicate 'messages' from a writer to a reader via a natural language text medium. Therefore, the study must regard information organisation in terms of its association with information transfer. That is, a structure which reflects any information pertaining to the text with which it is associated, cannot stand alone without some relationship to a theory of how information is transferred from one human to another by way of natural language text. An integrated theory of how information is trans-ferred from text within a communication system must be built. I will propose a writer-text-reader communication system to illustrate how information which is conceptually beneath the written word is transferred from text to readers. The use of the term beneath when referring to the relationship of information structures to the surface or syntactic structures of text, draws in part on the deep structures theory of CHOMSKY (1965). That is, semantic structures which reflect some inference which is represented by the surface syntax of the written natural language. The notion of an information structure is not synonomous with a semantic structure as such though. Information structures, as envisaged by this research, contain elements of both meaning (in the semantic sense) and organisation, in the syntactic sense. Conceptually, information structures are beneath the surface of the written language

of the text because they reflect more than the words and sentences that are merely ordered on the paper. That 'order' and the meaning of the words within the text together determine the organisation of the information in text which is not directly obvious from just the syntactic organisation of it. We require semantic labels and rules for applying them to the text to adequately produce a representation of individual authors' arguments. Although written natural language itself may be linear, semantic (including information) structures may actually appear non-linear. Demonstrating this point is central to the research because the fact that arguments may be presented non-linearly in text and yet conceptualised linearly by readers, gives us cause to examine the process of information transfer which takes place between a text and its readers.

In Chapter 2 I discuss how other disciplines are applying structural analysis (particularly with the aid of computers) to natural language text. I would argue that because these other disciplines are carrying out such analysis, even if their aims are to produce different forms of output, that is a very good reason for Information Science to be studying the nature of the information within the structures. As I said before, the study of information and its organisation is the responsibility of Information Science and the need for some results within the discipline is apparent. HUTCHINS (1977) outlined in some detail the problems of establishing subject 'aboutness' in text and he showed up the deficiency of a strong theoretical approach in document analysis. CLEMENT-DAVIES (1978) referred to Hutchins' work in his own evaluation of the problem and said, "The moral which may well be drawn from Hutchins' suggestive paper seems to be this: any effective algorithm to establish a document's topic by non-statistical means will have to be driven by an artificial intelligence of scarcely forseeable power". I do not propose a solution to this challenge which is quite as dramatic as the prognosis given, but I hope to show that some constructive work is being undertaken to aid the deficiencies of document analysis that are widely appreciated. It is as well to mention here that the information structure chosen to demonstrate information organisation and transfer does not in fact refer to subject organisation. The inferences from both Hutchins and Clement-Davies that some fundamental assumptions and explanations concerning information organisation are lacking, is the point to be drawn from their discussions so far as this thesis is

concerned. In my view, we can only attempt to solve the problems of sophisticated techniques for text analysis when we have coped with the deficiencies in terminology and research methodology. It is to this most fundamental of problems in Information Science which this thesis is primarily addressed.

There is a group of individuals who carry out research in a field known as the sociology of science. SMALL (1978) has examined the citations of authors of social science literature in an attempt to determine, "...the particular idea the citing author is associating with the cited documents ... the document is viewed as symbolic of the idea expressed in the text." This work seems to attempt to establish relationships between LEACH'S (1976) work which shows the logic by which symbols are connected in natural language communication and KUHN'S (1970) theory of conceptual structures in scientific communication which refers to the nature of citations of writers who are better or less qualified in the subject which they are writing. That is, the work of renowned authors is more likely to be cited by writers with less experience than those with considerably more. I presume that this is where the sociology of what is in my view essentially a bibliometric exercise is found. Another researcher in this field of the sociology of science SPIEGAL-ROSING (1977), writes about the introduction of a new science journal and assesses the 'image' which it projects to its readers and what is communicated about this 'image' back from the readers to the editors. In this case, as in the one before, I can see similiarities with my work because the topic refers to conceptual information processing and cognitive processes. I hope however, that the distinction between their work and what I am doing can be seen. I am actually attempting to describe what the information structure I have chosen to investigate 'looks like' and how this information is realised by readers. The comparison between my work and theirs ends with our common use of philosophical notions about existing phenomena and the study of conceptual information processing.

My research comes within an area of Information Science which has come to be labelled cognitive information processing. It could just as easily be called semantic information processing, or even just text processing. The point is, that Information Science remains dominated by problems of information retrieval and for the most part concentrates on citation data from documents rather than the text itself. Where

this generalisation is not so true is in the study of natural language problems, but even so the work to date has been of a more linguistic than cognitive nature. To a large extent the areas of interest overlap and it is difficult to differentiate between say linguistic and cognitive problems when we are dealing with semantics in natural language text. The problems of semantic labels might be of a linguistic nature but once we begin referring to meaning we must at least consider artificial intelligence - and that leads us into a whole host of other problems such as representing knowledge and learning processes. What I have tried to do in this research is to bring together the work of other Information Scientists who have examined problems like mine and integrate them to form a solid theoretical base for my own ideas. Later it will be seen that various aspects of my model for information transfer which has emerged from the research, can be related to at least four individuals' work in this area of Information Science.

1.3 ASSUMPTION UNDERLYING THIS RESEARCH

In short the basic assumption of this project, is that there is an ideal or an intuitive notion of a conventional format for the presentation of arguments in science text. This assumption is founded partly on observational evidence and partly on the replies of writers and readers of science text to questions during individual interviews. The two questions put to these individuals were:

(1) 'Do you think that there is a conventional format for the presentation of empirical arguments in science text?'

(2) 'If so, what do you think this format looks like?'

The consensus of opinion was that there was a conventional format and that it followed a three-phase progression from introductory statements of hypothesis or aim - to statements of data for the argument and method, experiments and results - to statements of conclusion.

In SALLIS (1978) I outlined my reasons for assuming the existence of a conventional format for presenting and representing empirical argument. In this paper I described the format as consisting of three 'phases' and of having concept descriptors for individual statements in text within each of these phases. These 'concept descriptions' are now better thought of as semantic labels which describe the type of individual statements within a notional format for empirical arguments. This essentially philosophical notion was identified by various means. For instance, in an informal interview with post-graduate science students, I asked them to write down what they thought was the form of argument presentation verbally and in text. They all produced representations which were linear and took the general form of what I had assumed to be a conventional format for arguments. Other researchers (see for example, BELKIN (1977) p. 123) have suggested that science text takes the general form of a problem description followed by a solution description, the solution itself and a description of results.

Jean-louis LAURIERE (1978) , describes the "common informational process ..." (p.32) as being thus:

(1) real life environment → (2) statement of a problem →
(3) algorithm of solution → (4) computer resolution.

He is concerned in this paper with constructing a language and a program for stating and solving combinatorial problems, which is why he ends with '(4) computer resolution'. It seems though, using his assumption about the common informational process, that he has a notion of the format used to represent an argument, leading from a statement of the problem up to the resolution of it. Later in his paper, LAURIERE gives his " ... formulation of the best hypothesis ..." which is an attempt to generalise three kinds of solution to any problem. That is, to any problem one can always assign,

1 : a feasible solution; or
2 : an approximate solution; or
3 : the optimum solution.

Although this is an attempt to define more closely the nature of the category solution, rather than add to the notion of a conventional

format for arguments, it does indicate the general use of semantic surrogates for different aspects of an argument. Whereas I might assign an organisational label to a statement of solution, Lauriere would assign one of the three categories above. This is just a difference in purpose.

In a recent paper for the Philosophy of the Social Science, ANDERSON (1978), gave some organisational features for the production of a plausible text. In this paper Anderson describes the existence of semantic categories in text which denote their plausibility. Amongst references to conversational analysis between groups within a given population to show presentation differences and terms such as 'indexicality' of text, he mentions semantic categories which indicate the organisation of authors' arguments. These categories are not the ones I am using for my purposes though. Anderson is checking thought sequence and cultural variations in originators and recipients. His overall conclusions are supportive of my assumptions though.

In all cases of this previous work, the representation of arguments is linear and generally in accordance with the three-phase format I have proposed. That is, introduction (hypothesis or aim), followed by method (evidence, citations, results), followed by conclusions. I repeat that this format is the product of a philosophical notion but intuitively and pragmatically to some extent it seems to be supported by others.

As a final point here I would mention that in my view arguments consist of a number of propositions, which in turn are represented by statements of fact; either true or false - see HUGUES and LONDEY (1965) and WITTGENSTEIN (1967) for one of three major philosophical opinions where this view of mine resides. In the analysis of text which follows later, statements are the semantic 'units' which are classified within the conventional format. In practical (or syntactic) terms, statements often seem to be synonomous with sentences, although obviously one sentence can for instance contain two statements.

Other observational evidence can be found which suggests that there is an ideal of the ways arguments are presented in science text. Out of 150 texts which I examined, 131 of them had section or paragraph

headings which follow the three-phase progression given above - that is 87.75%. Even if the individual texts are presented for publication in this way because of editorial policy, an ideal of how arguments should be presented is suggested. The above evidence, along with my own intuitive notions and the philosophical descriptions of well-informed arguments (see HUGUES and LONDEY (1965)), is the basis for my assumption that there is an ideal of a conventional format for empirical argument in science text.

1.4 AIMS AND PURPOSE OF THIS RESEARCH

The first aim of this project was to develop an analytical tool to be used in the recognition of individual statements in text. Having based my previously stated assumption largely on the observational evidence that 87.75% of my sample had section or paragraph headings which follow the ideal notion of a conventional format, I wanted to see whether or not individual statements which were grouped within these headings could be classified using the same label. In a wider sense. I was interested to see what patterns of argument presentation existed. Secondly, I was curious to see how readers interpret the organisation of authors' arguments and whether they adhered to the ideal of a conventional format in their production of summaries of arguments from whole documents.

The purpose of collecting data from the endeavours just given, was to see whether or not there is a theoretical justification for designing a system which produces highly structured summaries of documents according to the ideal notion of a conventional format for the presentation of empirical arguments. The motivation underlying this purpose is to provide some better understanding of the actual process of writing and reading science texts to those who are concerned with producing standards for report writing in science and technology. These broad endeavours obviously encompass several issues which are at the 'philosophical core' of the research.

Implicit in what has already been said, is the view that Information Science requires a great deal of work to be carried out in the area of

cognitive information processing and concept analysis of text. Coupled with this is a need to define our problem areas in this field and develop a methodology for investigating and solving them. Much of the initial work in this research therefore, was concerned with examining the fundamental problems of terminology and concept definition. Although this research is primarily concerned with developing a theory of information transfer, I was very anxious to bring some clarity to terms and concepts which are being used in the study of problems in Information Science. I refer particularly here to my description of the information structure itself. Generally speaking this structure has the facets which are intrinsic to any information structure. First there is the conceptual information which is, I assume, the 'meaning' of the argument in text. Second is the organisational or meta-information referred to earlier. The concept of a two-faceted information structure has been proposed before by SHREIDER (1974) and I hope to show in this thesis how I have used his previous work to develop the theory which I later describe. By using the work of others in this way I hope to contribute towards the development of a unified methodology for the study of problems in Information Science; particularly conceptual information processing and the relationship of cognitive processes to the field.

In discussing the nature of problems within Information Science, some of the overall aims of this research will be stated. The problem under review is in effect, one of definitions. That is, trying to identify which characteristics of text and information transfer are appropriate to the general problem of the organisation of information. As a vehicle for communicating thoughts, concepts, ideas or facts, natural language can in the linguistic sense, be thought of as a two-faceted entity. On the surface we see the written representation of the language and its syntax. Beneath the surface in what CHOMSKY (1965) called 'deep structures', we have the meaning or the semantics of the language. Information structures must, as has been mentioned before, fall simultaneously into both of these categories to some extent. Therefore, some of what this research hopes to achieve, is a clearer view of the nature of information structures and what they comprise. In so doing, I do not propose to survey the many types of semantic-type structures that can be generated from say subject or thematic analysis of text. That would detract from the central issue of the theory being

described. Some appreciation of the different types of forms of information will emerge from the discussion throughout the work though, especially in Chapter 2. PROPP (1968) in his classic thesis for morphological analysis of the Russian folk-tale, has demonstrated how structures which reflect the characterisation and theme of those stories can be generated. Similarly, RUMELHART (1975) and later THORNDYKE (1977) have produced 'story grammars' which can be used to segment text in terms of such descriptors as events. They have used these structures to aid in the evaluation of what constitutes memory organisation and how semantic-type information from text determines the storing of stories in human memory so that they may be recalled in the form that they were initially presented in text. SCHANK (1973) and WILKS (1976) have both developed theories which represent work in Artificial Intelligence concerning the nature and interpretation of natural language text in terms of its meaning. All of those investigations relate to problems within different disciplines, but in a way the problems themselves are 'applications' of the core issues which define each discipline. If these disciplines are concerned with other than the nature of information itself, then the study of information and its organisational properties is the problem for Information Science.

The grammar which is described later, is used to label statement-types in text and classify them within different phases of the author's argument. As mentioned earlier, this grammar consists of meta-informational descriptors and rules for applying them to statements in text. To be generally useful in an operational system, say for statement-type classification in a computer-based language translation system, the accurate identification of statement-types would probably have to be in the region of 85-90%. Therefore, any procedural rules which are developed to carry out this kind of text analysis algorithmically would have to cope with a multitude of linguistic and psychological variables and constraints. Resolving ambiguities in the English Language alone is a major problem. In Chapter 2 I have outlined some text processing systems and techniques. I have also shown that computer analysis of text to produce the kind of output required here has not yet been achieved. Some work, (see for instance, KLEIN (1962)), has been carried out to produce paraphrased summaries of text, but without using a semantic grammar approach. The emphasis has mostly been on lists of 'stop words' either functional (like 'Run' for verbs) or subject oriented like words from a thesaurus of subject terms.

I assume that human readers have the ability to cope with such problems as ambiguity when conducting statement-type recognition during their reading of text. I also feel that computer analysis of text is important for my work because in writing computer programs to carry out this kind of analysis (or at least trying to write adequate programs), we are forced to deal precisely and logically with problems which arise during the formulation of any rules. Whether computers are used or not, we need to have some notions of the rules which determine the input and output to and from the kind of structures I am investigating.

1.5 METHODOLOGY

I have begun with a premise that those who are concerned with the organisation and representation of information, need a better understanding of the process which exists for communicating arguments in science documents. I presume that arguments themselves are the semantic vehicle used by writers to communicate messages to readers. After some observational and intuitive evidence, I have proposed an assumption that there exists an ideal of a conventional format for the presentation of authors' arguments in text.

One of the greatest problems when studying any cognitive process (which writing and reading are), is to isolate and control variables such as memory, subject knowledge and natural language skills, when attempting to conduct any experiment which involves these factors. Human nature being what it is, individuals can give different answers to similar questions from one moment to another anyway. I do not therefore, pretend that any experimentation I conduct or results I produce are generally representive of humanity. All I would infer from my results is that they have been obtained from random samples of humans and literature and therefore, have some significance in relation to the phenomenon of information transfer from writers - to text - to readers.

A survey of existing research and problems associated with the analyses which I have done is presented in Chapter 2. This survey is I think, representative of current work and points to some general and often intractable common problems. Wherever possible I have defined terms and concepts in the light of my work and Information Science.

This is partly because I believe that Information Science requires to develop its own methodology for stating and solving problems, given that the 'scientific method' is not always appropriate to individual investigations and that the social sciences are fraught with imprecise and sometimes ill-defined terms and definitions.

As previously stated, one of the aims of this project was to construct a set of semantic descriptors which reflect the ideal of the notional convention for presenting arguments in science text. This set, (given in Chapter 3), was developed mostly from the consensus of opinion gathered from scientists who were interviewed at the outset of the project. An attempt was made to develop algorithmic rules to use in the analysis of text so that the set of descriptors could be applied to individual statements. The result of this endeavour and a description of analytical methods can also be seen in Chapter 3. Emerging from the development of this set of labels and rules (later called a grammar), was the description of an information structure where the labels became organisational elements and statements in text became data within categories classified by these elements. In effect what was produced was a meta-information structure, because the organisational elements refer to the format or presentation of the conceptual information, or the 'message' of the author's argument. Meta-information is discussed in Chapter 3.

The first stage of 'experimentation' was to classify individual statements within section-headings to see whether or not the statements were of the same type as the section-heading. The result of this testing can be found in Chapter 4. Text with no headings were also analysed. A large group of individuals then analysed one text so that I could check for patterns in their analyses and classifications. The results from these are also to be found in Chapter 4. A small sample from the Humanities (Philosophy) was analysed for comparative purposes. I then had individuals read a text and produce a summary of what they considered to be the author's argument. Their summaries were analysed and compared with the initial analysis of the text they had read. The result of this experiment which was repeated with other text and different individuals, was that although the whole text may have an argument organisation which does not conform to the conventional format, (even if its section-headings do), the summaries

of that text produced by readers, do follow the ideal format. These summaries were produced by individuals who had no notion of what my research entailed when they participated in the experiments.

My interpretation of the results obtained from the experiments mentioned above is given in Chapter 5. The outcome has been a model of the process of information transfer from writer - to text - to reader and a description of some implications for the theory which has been proposed. A computer system for producing highly structured summaries of documents based on the conventional format (my set of semantic descriptors) is outlined and demonstrated in Chapter 5, followed by my conclusions and thoughts for future work in Chapter 6.

The accent of this research has been on interpreting what results I have obtained from my text analysis of my own and others. I have not tried to normalise events or propose a generalised theory for all occurrences of the process I am describing.

My early work with information structures involved a good deal of background study in the theory of directed graphs particularly, and with forms of structural representation. For instance, trees, networks (or plex structures as MARTIN (1977)calls them), and the nature of the elements and their relationships within these structures. This work led me to the conclusion that there are two kinds of elements in information structures. These are the organisational or descriptive elements (which I later called meta-information from SHREIDER (1974)), which categorise the elements of conceptual information. These latter elements may be merely subject terms, but the relationship between any number of them signified by the organisational elements, reflects the information as a whole.

The next aspect of the research involved studying the dynamics of natural language and the various structures and analyses which are associated with it. This means a good deal of background work in linguistic theory, particularly computational linguistics where more formal models and rules have been developed for text analysis. Inevitably the work of cognitive psychologists and researchers within Artificial Intelligence made some contribution to my work. This is particularly true of that part of this research which refers to the formulation of

a grammar for analysing text in terms of the author's argument organisation. Computer Science and my own work with formal language in that field also contributed greatly to the theory that evolved. In most cases convention and intuition were the only criteria for developing the necessary aspects of the theory which developed. Wherever possible these philosophical notions have been examined and demonstrated during the thesis.

Formulating questions and answers throughout the research has proved a valuable aspect of the methodology because often I found that incidental issues arose which came to have some direct or indirect bearing on the theory proper. The type of questions asked have been, "Is there a structure inherent in text which can be identified, described and actually extracted from text, in such precise terms that computer programs could be written to do it?" To answer this question, we must investigate and define 'structure', 'information structure' and 'information' itself. This naturally leads on into the subject of communication systems and information transfer. Another question asked was, "Is it at all desirable that there should be a definite structure within which all writers of science journal articles (for instance), organise their arguments in text?" We can show that a structure does exist, but is this structure really desirable? This question raises issues of conventional format for writers and the information needs of readers. It is in attempting to answer such questions as these that the practical possibilities of any theory that is developed can be seen. If we are endeavouring to model something that actually exists, then the practice has already been established and we are only concerned with the theory underlying it. Some 'spin-offs' from the investigations do show the way to other systems though, outside of the practice that is being modelled. For instance, reliable statement-type recognition could lead to the useful indexing of documents in terms of their relationship with certain aspects of a subject state-of-knowledge. This may lead onto an ability to automatically paraphrase text or even produce abstracts of documents by computer. There are perhaps, consequences for information storage and retrieval systems in the interrogation of documents to establish aspects of author's arguments in those documents. These are peripheral to the general problem investigation, but issues which I feel have some bearing on my work. Some mention is made of these other issues at the end of the thesis where I discuss further work.

Unlike the physical sciences where an established theory or test can be repeated indefinitely to produce the same result, the theory which I am proposing in this thesis relates to the most fallible and unpredictable phenomenon on earth. That is, the human intellect and one human's interpretation of another human's ideas and written expression of those ideas. The very vehicle for communicating information and knowledge which I am studying, (natural language), is a human creation and can only attempt to symbolically represent concepts or happenings in the 'real' or physical world. Natural language is not itself a physical phenomenon, nor are the conceptualised results of human interpretation of any 'messages' in text. Therefore, to generalise any theory of information transfer to all human behaviour based on data or results obtained from any sample of humans is fallacious. The same sample of humans might produce different results on another occasion anyway. The aim of all experimentation in this research is to provide some credibility and substance for my speculative theory which I say <u>may</u> or <u>sometimes</u> is the case. I cannot and do not suggest a generalisation of the theory to all human behaviour. As I pointed out at the beginning of this chapter, I am taking an interpretative approach rather than a normative approach to this research. I am most concerned with interpreting what people do given a general and conventional situation, rather than trying to generalise my results to all humans who might be in the same situation.

1.6 <u>SUMMARY</u>

This chapter is intended as both a rationale or justification for investigating the thesis topic within Information Science, and also as an overview of the research itself. An assumption that writers and readers of science text have an <u>ideal</u> of a conventional format for the presentation of empirical arguments is made and evidence for that assumption is given. The main aim of the research is shown to be the development of a linguistic 'tool' for use in text analysis so that the organisation of authors' arguments can be represented in some structural way. This is to compare the practice of argument presentation with the notional ideal just mentioned. The ultimate aim leading on from the results of this text analysis, is the production of a theory for the process of information transfer from writer - to text - to reader. The

purpose of such a theory is to support a case for the production of highly structured document summaries based on the conventional format for arguments. A brief overview of what had been done in the research is given here to indicate the methods used during the investigation.

CHAPTER 2

NATURAL LANGUAGE TEXT PROCESSING

2.1 OVERVIEW OF THIS CHAPTER

The difficulties of surveying the topic of natural language text processing are two-fold. First, the definition of what this topic extends to creates difficulties because it covers both manual and computer analysis of text and the processing of different documents for various reasons. It could be the subject indexing of documents, the analysis of meaning in text, thematic analysis, or as in my case the discovery of semantic categories which show the organisational properties of text. Second, although there are many references to work in this general area, much of what has been produced refers to theoretical systems or at best simulations and models. Systems which are fully operational appear to have serious conceptual limitations and in some cases clearly do not do what is intended to be done. This all makes the evaluation of existing systems very difficult.

This chapter begins by pointing out some of the natural language constraints which are placed upon any theory or system for text processing. The term text processing and a discussion of semantics for Information Science is given, followed by a brief survey of some of the major natural language text processing systems which are operational. It is intended that this chapter should first and foremost be a base from which to appreciate the complexities involved in the formulation of a grammar such as I have had to produce for the description of authors' argument organisations in science text. Machine translation (MT) systems are discussed in some detail because they encounter the problems of syntactic and semantic description and concept organisation in text, much the same as I have to cope with statement-type classification. I am not trying to translate the natural language but the categorisation

of semantic entities is a common problem for designers of M.T. systems as well as for me. Generally speaking, it is useful to study computer analysis of text because the algorithmic techniques used in such systems attempt to define precisely rules which can be applied in other forms of text analysis. As this chapter indicates, the immediate future for M.T. systems does not appear to offer any significant alteration in the methods or techniques used.

2.2 PROBLEMS OF NATURAL LANGUAGE TEXT ANALYSIS

The literature of computational linguistics, (this includes work by researchers in Artificial Intelligence, Computer Science, Psychology and Information Science which relates to problems in computational linguistics), seems to project a common message. That is, natural language contains certain semantic properties which can be exploited in an endeavour to illustrate various inversions of text and verbal discourse. By 'inversion' here, I mean views of representations of text. In generating any such output, the data (the text), should not be permuted or altered in any way to facilitate processing. If we change the data by coding or pre-editing it, we are removing those properties in text which exist in 'real' situations. The message that comes from the literature referred to above is that we must deal with natural language text as it is written and any significant alteration of it invalidates the analysis being carried out. Conversion of text to some classification scheme before analysis would enable more reliable processing to occur of course, because ambiguities for example could be eliminated. We must look within the written language for syntactic and semantic 'clues' as to the nature of, in my case, statement-types. The conclusion reached so far, is that it is immensely difficult to accurately identify and implement on a computer, the intellectual experiences and interpretative skills of human readers.

For a comprehensive study of identifying the performative verb in human discourse through the medium of natural language, see FISHER (1977), who developed a computer system to analyse input sentences for such 'type recognition'. This is a good example of an attempt at statement recognition. GOSHAWKE (1976) is an example of someone who has developed a computer system for translating one natural language into another by means of pre-classifying words into a numeric scheme. Goshawke's work is conceptually unsound in terms of the discussion just presented,

because he converts all natural language to arbitrary codes (numbers) which makes his system most inflexible and only operates on short and extremely simple sentences. The main problem which faces computational linguists in particular, is how to develop procedures which will analyse natural language text to produce a variety of inversions or translations of an acceptable quality, without disturbing the intrinsic nature of the writing.

I do not propose to venture too deeply into the linguistic theories of syntax and semantics, but some examples of both are useful for differentiating between the two when it comes to the design of parsing algorithms later in the thesis. As a very fundamental example, if we were carrying out word recognition analysis to establish the frequency of the word 'jump' in a particular text, we would be conducting what is essentially a syntactic analysis of the text. If we were to then establish by some means that the word 'jump' meant say, 'to leave the surface on which one is standing and be suspended in space for a finite period of time before landing on the same or a different surface', we would have conducted some semantic analysis.

WEINGARTEN (1973) illustrated the ambiguity of the English Language with the two examples given below. He used the syntactic labels attributed to English sentences and constructed two different organisations for the same sentence. The first example of the sentence 'They are flying planes' indicates that some beings are piloting planes. The second example of the same sentence indicates an observation that some objects happen to be flying planes. The other useful aspect of these examples is that they demonstrate just how important the syntactic organisation of text is for the correct or intended semantic inference to be communicated. Following these two examples there appears a further syntactic organisation to show how the formal language of mathematics can be demonstrated on the expression (A + B) x C. In this language there can be no ambiguity. See KAY (1967) for examples of some early work with semantic reduction to eliminate ambiguity of text.

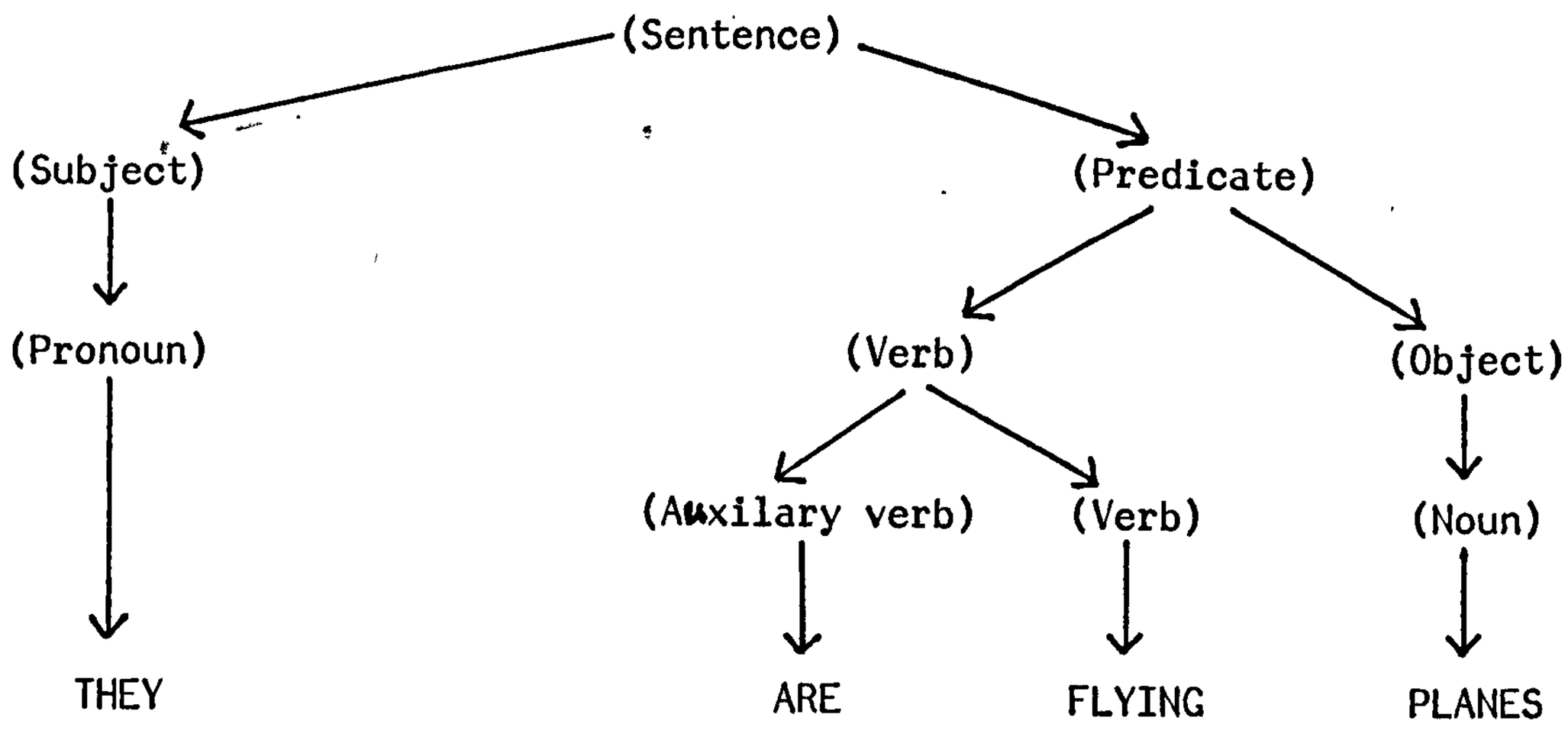
One way of overcoming the vagaries of ambiguity and such like in the English Language, is to build a set of meta-linguistic labels which can be used to describe various semantic aspects of text without relying on the grammar rules that are associated with that language. Such a set of labels can be referred to as a pre-defined, restricted grammar. Pre-defined, because we state exactly what can be referenced by the labels

FIGURE I

SHOWING TWO DIFFERENT SYNTACTIC STRUCTURES

FOR THE SAME ENGLISH LANGUAGE SENTENCE

(a)



(b)

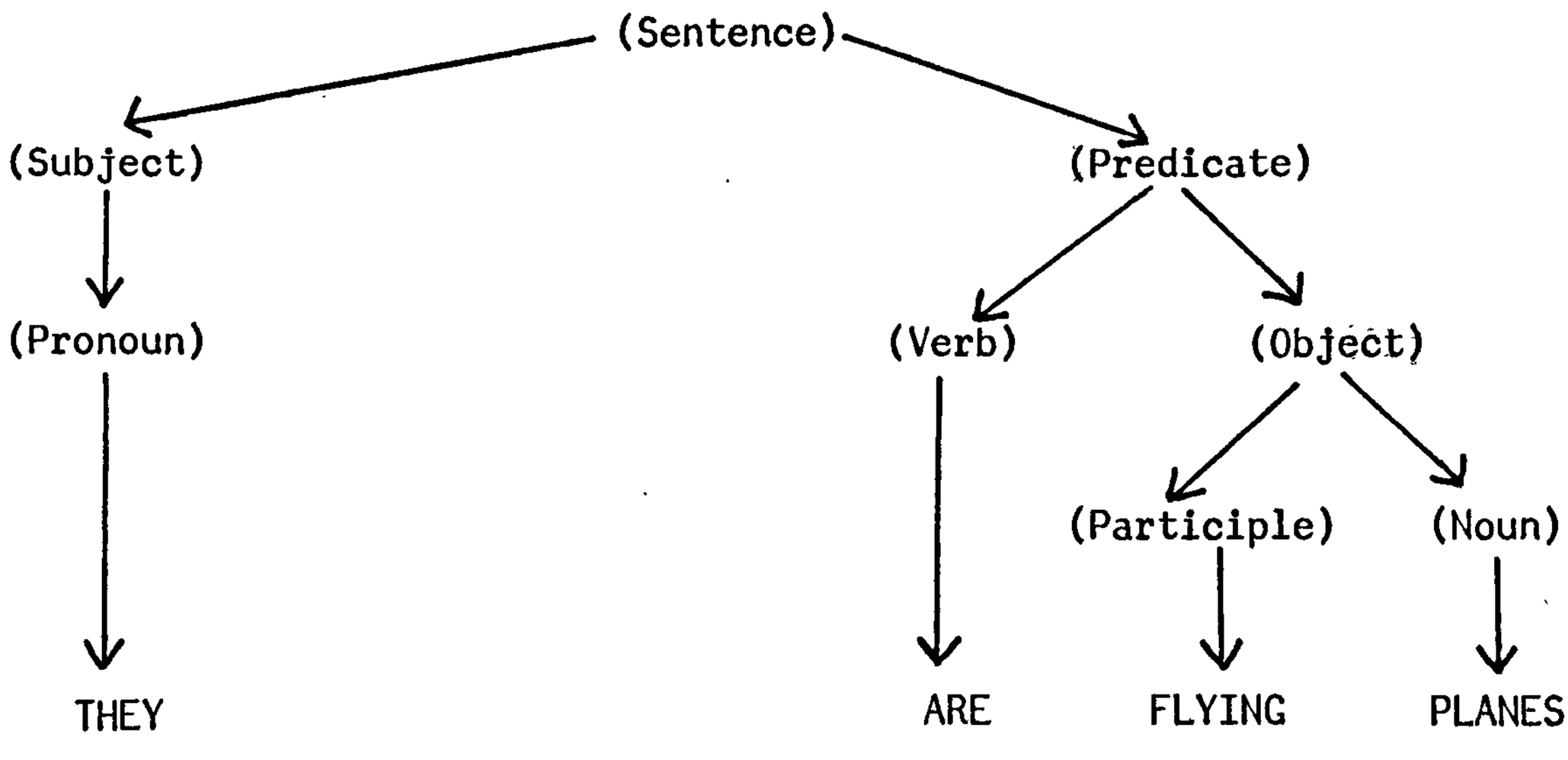
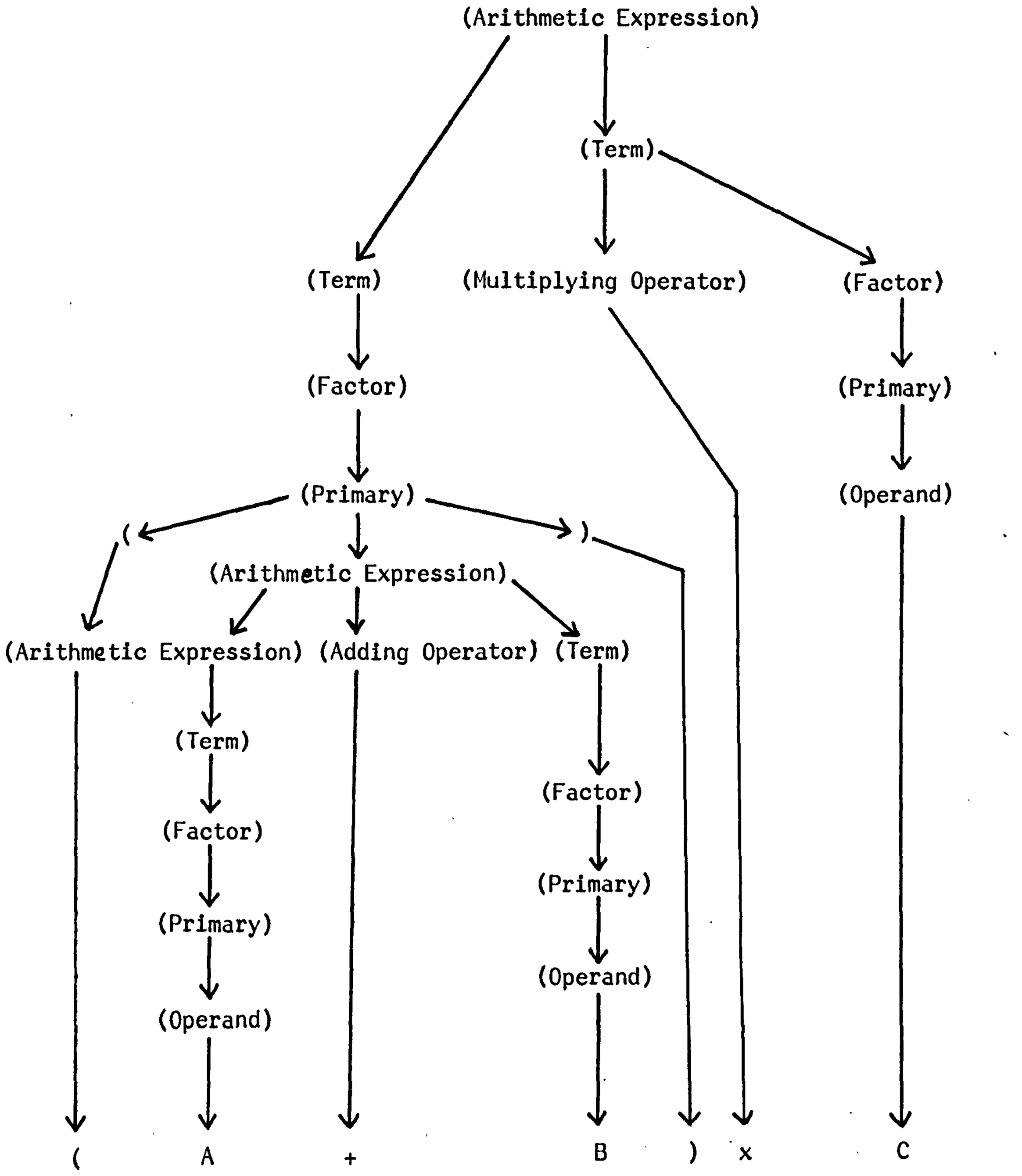


FIGURE 2

MATHEMATICAL SYNTAX STRUCTURE FOR THE EXPRESSION (A+B)xC



and restricted because the grammar could not apply to all classes of written language. In our case we are only referring to the organisation of arguments in science text. This means we can refer to aspects of the text which indicate various phases of the conceptual argument of the writer. If we introduce a grammar such as this we will need to contribute to the idea of controlled vocabularies in as much as the contents of text which will fall into the categories of the grammar will need to be stated and defined before analysis can take place.

Experiments by SAGER (1975) have shown that sublanguages exist within natural language text and that grammars such as I propose for this research can represent "... the contents of natural language science texts without resorting to surrogates or a priori semantic categories ...". As stated though, these grammars are within science texts and highly specialised fields where categorisation of discrete aspects of text are more precise than in the humanities. Her earlier work (SAGER 1972), was entitled Syntactic formatting of science information, which was even more obviously relevant to my work because she discussed the potential of sub-language grammars for formatting science text.

Sublanguages have been the province of linguists for some years and have been acknowledged as existing within highly specialised literature, particularly science text. A notable contribution came with the mathematical proof of HARRIS (1968) for various linguistic structures in text. With regard to sub-languages, he believed that the transformation of a sentence in one discipline to a similar sentence in the same discipline, was possible through the existence of sub-languages. He regarded 'whole language' as consisting of sentences and called those sentences sets. Sub-sets were parts of sentences thus, they constituted a number of sub-languages. So, S(si,sii...), where S = sentence or whole set or whole language, and si... = a number of parts of sentences or sub-sets or sub-languages. Although Harris does not specifically say that this is only sometimes the case, I personally do not think it should always be thought the case. We are only considering that text which contains a precise and highly specialised subject terminology. The main inference I draw from his work, is that semantically at least, there exist structures in science text which can stand alone as transferable information, beyond the

whole text representation. This being so, Harris offers two important contributions to my work. First, the sub-sets to which he refers are essentially phrases containing those discipline-oriented words which could be taken out of the whole text context and yet the meaning of the text could still be understood. That is, a much less intelligible text could be created from the original, which would still be as intelligible to a subject-aware reader. KITTREDGE (1978) demonstrates this point using seven different types of every day communication where he believes sub-languages exist and yet we still comprehend the full message of the author. Amongst his examples are weather forecasts, newspaper editorials, cook books and home appliance instruction books. He gives examples like the following instruction. 'Put driver in screw and rotate left for three turns.' This sentence leaves out at least two occurrences of the word 'THE'. Thus, Kittredge says, (and Harris too), a whole language grammar which would insist on the use of the word 'THE', is inappropriate for communications such as these. The creation of sub-grammars must be undertaken to accurately define the language being used. This is the first point from the work of Harris and the others in sub-language research which is most important for my thesis. In my view, the creation of an appropriate set of descriptors to reflect the organisation of some information in text is essential, but this endeavour is necessarily coupled with the task of revealing the existence of the organisational properties in the first place. The second useful aspect of Harris's work assists here too. As I pointed out above, the parts of sentences referred to are phrases. I shall show in the next chapter how phrases like, 'The aim of this paper is ...", are perhaps the most reliable indicators of statement-type in text and ultimately of argument organisation. There are therefore, some obvious parallels between structural components for theories of sub-language and meta-information, the main one being the reliance on semantic indicators in written text.

A great deal of the work that has taken place in the analysis of meaning has been within specialised subject fields. For example, WILKS (1973) proposed the use of paraphrasing, (including jargon and formal technical terms), as a means by which information could be represented out of context for subject-areas which made 'heavy use'

of precise and meaningful terms. WOLFF (1976) and other cognitive psychologists have carried text analysis in specialised subject areas which they call elemental segmentation. These studies operate on 'chunks' of text which are said to contain the essence of the meaning of the writer. Their techniques are designed to overcome problems of redundancy and ambiguity, but rely on principles of the semantic differential ('loaded words', or terms which by their very existence denote discrete concepts), for recognising relevant 'chunks' in text. When reduced to its most atomic state, such a theory merely means that the analytical process depends on matching words in text with pre-defined stored lists of 'function words', or words which mean something in a particular subject field. As we shall see during this chapter, this analytical process is still the most used and most viable means of recognising aspects of text. See HOLMES and WATSON (1976) for a discussion of the roles surface order and syntactic organisation play in semantic perception of the contents of sentences. Although proposed for use in psychology research, this work is of use when considering my assertions that text must be ordered in some structural semantic sense to be optimally informative.

I have mentioned the work of PROPP (1968) which established characteristics and themes in Russian folk tales. In Psychology, the way stories are interpreted and memorised is a subject for much discussion. RUMELHART (1975) developed a grammar which could be used to represent passages in prose. More recently, MANDLER and JOHNSON (1977) have adapted his work for use with children's stories and THORNDYKE (1977) uses yet another modification of it for his analysis of stories. His grammar is detailed below. I have included this here because it serves as a good comparison with the set of semantic descriptors which I have developed to produce organisational information from text. Most of this work refers to BARTLETT's (1932) work with memory retention for stories and how aspects of text are stored in human memory. In the grammar shown in Figure 3, Thorndyke has not included any specific rules for how the labels should be applied to text. He assumes, I imagine, that algorithmic or procedural rules are unable to be produced. Therefore, interpretation of the text and application of the labels is carried out pragmatically by humans.

FIGURE 3

THORNDYKE'S GRAMMAR RULES FOR SIMPLE STORIES

RULE NUMBER	RULES
1	Story → Setting + Theme + Plot + Resolution
2	Setting → Characters + Location + Time
3	Theme → (Event)* + Goal
4	Plot → Episode*
5	Episode → Subgoal + Attempt* + Outcome
6	Attempt → (Event* (Episode
7	Outcome → (Event* (State
8	Resolution → (Event* (State
9	Subgoal) Goal) → Desired State
10	Characters) Location) Time) → State

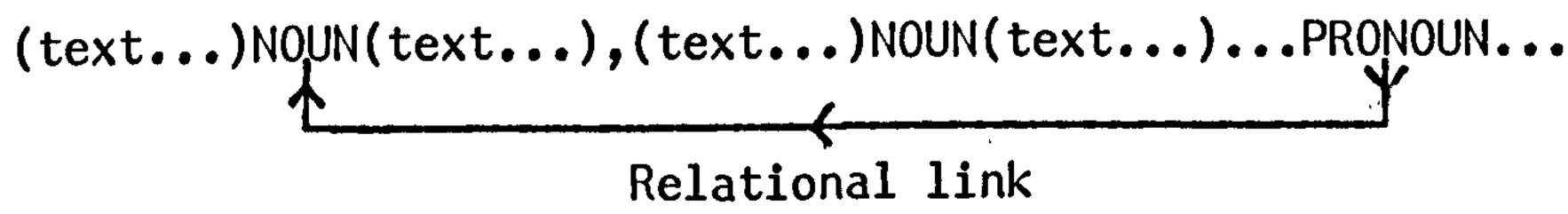
The Symbols: → Means production

\+ Means a sequential element

* Means repetition

Another approach to the problem of representing syntactic and semantic aspects of natural language text has been to use highly formal notation such as the Predicate Calculus to express the properties and operations of natural language. This is particularly evident in work relating to Phrase Structure Grammars (PSG's) - see BAR-HILLEL (1964) - and in attempts to formalise semantic categories in a similar fashion to the work in syntactic theory - see KEMPSON (1975). Had this approach been possible for this research, which I do not think it would because of its inherent inflexibility, the development of a much more involved meta-language would be necessary. In any event, such techniques are totally unwarranted for the problem to which this research is addressed. Highly symbolic notations are usually too far removed from natural language and mostly only represent the syntax of text; even if they do remove ambiguity thus making computer analysis easier. Formal languages such as mathematics, music and computer languages are self-descriptive. The use of Backus-Naur Form (BNF) to describe computer programming languages is useful because of its ability to simplify and disambiguate the syntax of these languages. As with the other algebraic notations, this is basically all it does. Good examples of the use and underlying theory of BNF can be found in WEINGARTEN (1973) and KNUTH (1972).

To further expose the difficulties of processing natural language data, let us now consider the formulation of just one, but very important rule. Structural elements, be they meta-linguistic labels or words from the text being analysed, are the nodal entities of our output organisation. In its most abstract form, the structure is defined by the rules which tell the elements where they are to reside in relation to one another. A good example of this is a rule which might handle anaphoric references. Let us say that this rule states that when a pronoun is encountered, a link is established between itself and the last encountered noun. In the case of two nouns appearing prior to the pronoun, where a comma also exists in the sentence, a pronoun will be linked to the noun before the comma. That is,



Obviously this particular link is not reciprocal but what does happen is that the pronoun becomes an information element in that it tells us

that the concept being communicated in the piece of text it is in, is directly related to that in which the 'linked to' noun is in. This rule is far too inflexible for actual use, but it does show the difficulties of establishing general-purpose algorithms for even the most fundamental problems of syntax parsing.

2.3 TEXT PROCESSING AND SEMANTICS FOR INFORMATION SCIENCE

The term text processing has, like many other aspects of the vocabulary of emerging disciplines such as Computer and Information Science, come to refer to a number of operations which are carried out on text of one form or another. In my opinion the term as used for Information Science can relate just as properly for the kind of analysis I am carrying out as to the editing function for which the term is often used in Computer Science. See SALLIS (1978a) for a discussion of this point. This does not mean that the various concepts and applications of the term by other individuals should just be ignored. Merely, that the use of the term for this research has a direct relationship with my concept of it for work in Information Science. In fact, the article referred to above refers to this research by way of example during discussion of the term.

The 'type' of text analysis carried out by inverting on the author's argument, can be seen in relation to other forms of concept analysis. In Figure 4 below, a ranked list of parsing strategies appears as four broad 'types'. This ranking shows the inversion of text on author's arguments, as occurring second in depth to meaning analysis - where depth is used in the Chomskian meaning, as discussed in Chapter 1.

FIGURE 4

RANKING OF TYPES OF TEXT ANALYSIS IN ORDER OF THEIR 'DEPTH' OF INVERSION

TYPE	EXPLANATION
Syntax parsing	To establish patterns in the data that are identifiable from the writing itself. For example, recognising words, sentences and punctuation. Keyword indexing and the greater part of machine translation (to date anyway) falls within this category.
Event parsing	Categories or types of events which are expressed in discourse (say events in a story) are classified during this kind of analysis. This is a semantic-type analysis - conceptually beneath the surface writing.
Argument parsing	Similar status to Event parsing but tending more towards the communication of 'ideas' or 'concepts' and therefore, the interpretation of meaning in the text.
Concept or Meaning parsing	Entirely semantic in objective. 'What is meant by such and such a phrase in text?' Very complex relationships between elements (both syntactic and semantic) in text have to be established in an endeavour to discover the meaning of what is written.

The ranking just given, is to demonstrate where the inversion of text on author's arguments is seen to reside in relation to other forms of text inversion. At the top of the rank we see parsing which is carried out to produce syntactic structures from text. For instance, the identification of nouns, verbs, conjunctions and so on. At the bottom of the rank appears the other linguistic category - semantic parsing. LEECH (1974) produced a summary of what he considers to be the seven types of meaning, within the field of Linguistics. His summary appears as Figure 5 below.

FIGURE 5

LEECH'S SEVEN TYPES OF MEANING

<table>
<tr><td colspan="2">1. CONCEPTUAL MEANING
or <u>sense</u></td><td>Logical, cognitive or denotative content.</td></tr>
<tr><td rowspan="5">ASSOCIATIVE MEANING</td><td>2. CONNOTATIVE MEANING</td><td>What is communicated by virtue of what language refers to.</td></tr>
<tr><td>3. STYLISTIC MEANING</td><td>What is communicated by the social circumstances of language use.</td></tr>
<tr><td>4. AFFECTIVE MEANING</td><td>What is communicated of the feelings and attitudes of the speaker/writer</td></tr>
<tr><td>5. REFLECTED MEANING</td><td>What is communicated through association with another sense of the same expression.</td></tr>
<tr><td>6. COLLOCATIVE MEANING</td><td>What is communicated through association with words which tend to occur in the environment of another word.</td></tr>
<tr><td colspan="2">7. THEMATIC MEANING</td><td>What is communicated by the way in which the message is organised in terms of order and emphasis.</td></tr>
</table>

<u>Thematic meaning</u> in Leech's interpretation, is the nearest comparison between one of his 'seven types' and the argument inversion I propose. This is a useful attempt at classifying types of meaning, particularly because it reflects the difficulties encountered by the compiler of such a summary when it comes to actually parsing text in order to invoke one or other of these classification terms. <u>Affective</u> meaning for instance, provokes all kinds of thought on how one might implement an automatic analysis of text to produce this type of inversion. We would have to code up a number of feelings and emotions and relate them to words in text to which they might refer. Complex rules for distinguishing between one emotion and another when words were in or out of context with a particular situation, would also need to be constructed. Take for example, the use of an oath (or swearing) in text. Such expressions can apply equally to anger and joy.

Although predominantly a linguistic study at this time, I can see that analyses to produce information structures along the lines of collocative meaning for example, will be useful to us. This type refers to the association of text words with similar words which have different meanings in other contexts. In attempting to match ambiguous terms during automatic information retrieval, procedures to produce text inversions which reflect such associations could be extremely useful.

In short, the identification of information and its organisation in text seems to me to be the immediate concern of Information Science, as I have mentioned before. A great deal of work in bibliometrics has produced results from which organisational information can be found. See for instance, BOTTLE and PERRY (1977) where they conducted subject analysis of titles from agricultural journal papers to examine the predominance and distribution of keywords in that subject area. The organisations that this work endeavours to produce relates specifically to information which tells us what the text being analysed is about, in terms of the subject area to which it refers. Similarly with TOCATLIAN (1970) which examines the information content of chemical paper titles and BUXTON and MEADOWS (1977) which attempts to measure variations in the information content of paper titles. I consider my work to be one level 'lower' than these other investigations, because I am trying to identify semantic categories and thereby information in the natural language itself, not just in the titles of text. In any event I am concerned with organisational rather than subject information.

2.4 A DESCRIPTION OF PREVIOUS WORK

Structural analysis of text (particularly by computer) is the subject for investigation by several fields other than Information Science. What follows is a brief discussion of some of these analyses. The preceding pages of this chapter already refer to some work of this nature, which I will not repeat here. SPARK-JONES and KAY (1977) gave a summary of Linguistics and Information Science in which they confirmed their earlier (1973) beliefs that information retrieval systems did not reflect much in the way of linguistic theory. I would make it clear that my analysis of text does not make claims to this body of theory either. I am concerned only with organisational information insofar as my grammar is concerned - although I obviously acknowledge the existence of conceptual information as forming part of the information in text too.

To begin this survey, I will first examine the developments and achievements of Machine Translation (MT) systems, because they have a direct bearing on my work in three respects. First, the study of computational linguistics generally, owes much of its background and theory to the early work in Machine Translation. It is therefore, important for us to note the methodologies employed in data organisation and algorithmic description of text processing techniques used in these systems for machine translation of natural language text, which have been developed. Second, the results of processing data for machine translation have encouraged researchers to suppose more definitely that conceptual structures exist in text and that these structures can be described and interpreted in text. Third, the matter of information needs by users, the organisation and manipulation of information to supply those needs, are foremost in the philosophy of machine translation. These systems are attempting to make available the written work of researchers in all natural languages, so that everyone may have the benefit of particular states-of-knowledge for a wide variety of subjects.

A good deal of recent research into machine translation is being conducted by groups into Cognitive Psychology and Artificial Intelligence. The latter group contains several psycho-linguists who have depended heavily on the technological and methodological developments of Computer Science over the past ten years or so. Research into the problems of natural language text translation has been going on since the early 1950s, so in many ways the subject has grown up with Computer Science, or at least with the computer industry generally. By 1963, when the United States Department of Commerce reported on developments in the area of machine translation, the future for the topic did not look very promising. The United States Government, particularly the Air Force, had been spending a large amount of money for the development of machine translation systems, but the cost of storage and processing time was high and the accuracy of the translations was low. The technology of the time was predicted to be unsuitable for any major improvement in the service offered for some long time in the future. Thus, many projects were abandoned and apart from a few dedicated individuals, intellectual and physical effort was diverted to other problems.

Two well known machine translation systems are worth discussing briefly here. The first is called SYSTRAN (for System Translation), and is a commercial product which has had some success both in the United States and more recently with the European Economic Commission.

The system (see TOMA (1977), for a description of it by the designer), uses massive dictionaries for both the source and target languages being processed at the time. It relies mostly on equivalences of terminology in both languages and as a result of this approach, has up to 45% post-editing problems. As WILKS (1977, 1977a) and others have pointed out, SYSTRAN would appear to be a very sophisticated re-write of some software written by SYSTRAN'S designer in 1962 - a software suite known as the 'Georgetown Programs'. This point is central to our discussion here. Nothing much has changed, conceptually at least, in the approach to machine translation, over the past fifteen years. In all fairness though, SYSTRAN is the largest operational machine translation system so far developed and it is working. It may not be one hundred percent accurate or infallible, but it is translating Russian to English, German to English and back again, French to English and back, and Spanish to English and back.

The Chinese University of Hong Kong also have an operational machine translation system known as CULT (Chinese University Language Translater). Although the level of accuracy appears high (up to about 80% - see LOU (1977)), still no radical change in concept has occurred for the design of the system. Dictionaries have still had to be built for term equivalence checking. In this system, students at the University coded Chinese pictagrams into machine-readable form, from which the dictionary was built. Like Goshawke's work, mentioned earlier, CULT translates into numbers before carrying out the 'natural language' translation. Mechanical rather than computer translation for the Chinese Language is referred to by MOSS (1978), who discusses research with a character encoding machine which can be used as fast as an electric typewriter to enter English into Chinese via a machine-readable form. The encoder converts the Chinese characters into a form recognisable by a computer, which then carries out the processing required. This way the reading of say Chinese then translation into English is avoided. This does not solve any of the problems encountered by those attempting straight translation of course, but it does provide evidence that some people are working on the problems from a different point-of-view. Moss is dealing with a pre-editing problem rather than a processing one.

Philosophy was another field to contribute to the methodology of machine translation problems. In light of the linguistic experiences

and theory which were developing from people like Chomsky in the late 1950s and early 1960s, philosophers began re-examining such phenomena as traditional belief systems. Some, like BAR-HILLEL (1964), have postulated complex logical explanations for structures and processes in natural language and provided exhaustive proofs using the Predicate Calculus. Psychologists began relating the storage of concepts in human memory with the various semantic structures of natural language. Many models (see particularly KEPPELL (1970) have been developed to represent this sort of cross-correlation between text (or discourse generally) and human memory. Much of the Chomskian philosophy continues to be used to develop theories using, for instance his concept of transformational grammar (1965) to develop new theories in psychology - see WOLFF (1976).

The common threads of all this work by researchers in several fields, were brought together in the late 1960s by a discipline which came to be called Artificial Intelligence. This field now covers a wide range of topics which concern investigating, modelling and attempting to emulate human behaviour - particularly cognitive behaviour. It is from this field that the most significant developments in the area of text understanding have arisen. Work by SCHANK (1971) and later COLBY (1973) is representative of contemporary research into semantic networks and what have come to be labelled meaning structures. WINOGRAD (1973) developed a computer program that learned about its environment during an interactive dialogue with a human user. NORMAN and RUMELHART (1975) together with a group of their colleagues have done a considerable amount of work on both sentence analysis and concept perception. BOBROW and WINOGRAD (1977) are now developing a special-purpose computer programming language called KRL (Knowledge Representation Language), which is being devised to state directly problems of knowledge organisation with a view to enabling the program to actually learn as it is processing data. Similar work being carried out in Germany (Sarrbruken COMSKI project - not yet documented), endeavours to produce a programming language for processing information generated from a knowledge base which is fed into the program as data.

In his recently published and very extensive book, BRUDERER (1978) surveys the existing systems and designs of systems for machine and machine-aided translation systems and data-bases and concludes that now, sixteen years after the United States Government withdrew its support

for the majority of the projects being worked on in that country, nothing much has changed in terms of methodology for solving the problems of natural language text translation; and there is nothing likely to happen for some long period of time. For a survey of trends and progress in the Soviet Union see KNOWLES (1979). The Soviets are most concerned with semantic categorisation and the representation of meaning - a trend we in the West are lacking progress in.

As part of the general text processing milieu, various language-types can be ranked according to their suitability for machine or machine-aided translation. As a final word on the topic of machine translation then, Figure 6 below lists a rank of languages from the very informal language which has a random grammar, to the very formal language which has a strict pre-defined grammatical structure and rules for its use.

FIGURE 6

LANGUAGE-TYPES RANKED FOR MACHINE-TRANSLATION SUITABILITY

Randomly generated language	-	random grammars/random words e.g.: 'Jabberwokky'.
Natural language	-	well defined grammar, but full of ambiguities when used in and out of context.
Abstracting language	-	same as Natural Language, except that the labels for parts of the text are different. Still subject to ambiguities in the language.
Indexing languages	-	more formal, in that they refer to actual words in the text.
Thesaurus language/ Dictionary language	-	much like Indexing Languages. Use of keywords makes the identification of actual words easier or more precise - transliteration.
Computer languages/ mathematics/music	-	highly formal grammars.

This ranking is not based upon any empirical testing, but is the result of a very methodical discussion with colleagues who have worked in the area of machine translation and information processing. It is useful for this research to note whereabouts the abstracting language comes in the ranking. It shows once again that although the meta-informational elements of the ensuing information structure from any analysis of text may be precise, the 'data' of the structure is still natural language itself. That is, statements from the text. Formal language translation is of course, a different problem from that of natural language translation. Many of the terms used (or vocabulary) are similar in both areas of research, but the problems of analysis differ considerably. Elaborate descriptions of phrase-structure grammars (PSG's), left-to-right parsing grammars, notably LR(k) where k is the element being scanned for in a left-to-right parse, and other such special-purpose grammars for translating computer programming languages can be found throughout the Computer Science literature - see particularly KNUTH (1971), BOOK (1978), and HUNT and SZYMANSKI (1978). Most courses in Computer Science now teach this area of the subject; usually in association with a course on language compiler design - see WEINGARTEN (1973) for an introduction to the topic.

The ability to disambiguate statements in both formal and informal languages, is one of the single most sought-after aspects of analytical procedures. In Appendix A I have included a computer program (written in ALGOL 60) which uses a grammatical structure based on the early work of CHOMSKY (1957), to show how sentences can be randomly generated using correct syntactic labels but containing semantically nonsensical words. The words are all regularly used in the English Language, but the program shows that even if they are strung together with syntactic accuracy, the eventual meaning of the sentences produced can be nonsense. Seemingly trivial experiments with natural language such as these can provide very useful information concerning the nature of the problems we are dealing with. The need for rules which establish semantic relationships between words in text is one of the major lessons learned from this exercise.

The subject of concept analysis to aid document indexing is not a new one. The classification of documents by their subject relationships has a long tradition of using schemes such as the Dewey Decimal Classification and based on the indexing theories of philosophers like

RANGANATHAN(1957). Recently, a computer-assisted indexing system called PRECIS (see AUSTIN(1974)), has been used by the British Library to construct lists of codes which represent indexing terms kept in a computer-readable thesaurus for each document entered in the British National Bibliography. Once again, although some of the conceptual foundations of the system seem inadequate to satisfy all of the problems encountered by the system, and certainly some indexers (see LANGRIDGE (1976)), PRECIS does fulfill the requirements of an automatic indexing system capable of assigning subject thesaurus terms to documents.

Automatic (or computer-assisted) abstracting of journal papers is another topic again. The French Textile Industry have a system called TITUS (Textile Industry Text Understanding System), which carries out some form of automatic abstracting. The computer does not read natural language text then construct an abstract of it though. Documents are indexed, then the terms are supplied to the computer system which contains a large number of pre-constructed sentences in the general form of an abstract. Once the terms are included with the appropriate sentence, an abstract is generated. Such a system can output abstracts (or at least summaries) of whole documents which are understandable by users of the system because once again the terms used are 'loaded' relative to their technical significance. The reader is interpreting the abstract or summary by recognising the concepts which are represented by the combination of subject-related words. This is not automatic indexing because the system does not carry out any linguistic processing beyond assigning phrases to index terms. The time when a computer can use semantic rules to produce linguistically sound summaries of whole text is still some way off. Given sufficiently flexible and comprehensive rules, a grammar like mine might be used as part of this kind of process. This idea is discussed further when I describe the grammar more fully in Chapter Three.

Most of the systems I have described here use stop lists of terms to extract corresponding words from text. The term 'stop list' is synonomous with 'go lists' as they are usually called in key-word indexing systems. Both terms appear in the literature to mean the same thing. That is, a list of subject-related words which can be used to match against words in text in order to produce indexes or concordances and the like. Lists such as this often suggest controlled vocabularies for the analyses they are being used for. Some work with dynamic updating for such lists, which are often

structured as thesauri, has been attempted. The successes have been minimal, but if it were possible to include new terms in the lists during processing, we could have more linguistically flexible systems. See LANCASTER (1968) and SALTON(1973) and SALTON, YANG and YU (1975) for discussions of this problem.

Using the techniques of matching words in documents with those lists just described offers the advantages of accurate and relatively fast index and concordance generation. Running texts against such lists for word frequency analysis can also provide useful information about the distribution and use of subject-terms within particular subjects. The Semitic language retrieval system KEDEMA (see ATTAR et al (1978)) uses an extremely complex keyword expansion technique to cope with morphological analysis for document retrieval in Israel. Suffix-stripping in English (using the first significant part of a term; say, 'describ' from describing), is useless for Hebrew. A complex lexical analysis of search terms and their corresponding document matches is necessary and this is done by synthesising all target terms until they conform to a known term. It is with these kinds of text analysis that the advantages of computer-assistance can best be appreciated. Computers are fast, capable of handling large amounts of data and carrying out the process of synthesis required in say the KEDEMA system, mentioned above.

I have no intention of detracting from the excellent work carried out in Artificial Intelligence which is exploring the problems of knowledge representations and the analysis of meaning in text. The complex structures of CHARNIAK(1976) or BOBROW and WINOGRAD(1977) for expressing learned facts for instance, provides us with many 'clues' to the nature of knowledge and the synthesis of new information on existing states-of-knowledge. The plain fact is though, that although several systems for analysing text algorithmically have been designed, very few have reached an operational stage of development. In my view, the progress will be made by those projects such as the FOCUS PROJECT (see DREIZIN(1979)), which has developed a set of linguistic descriptors and rules to carry out essentially semantic analysis of Hewbrew-sacred legends. Unfortunately, the actual analysis is as yet purely syntactic and will require a great deal more work in the association of concept terms with phrases in the whole text before true semantic representations can be produced. The researchers are not pre-editing the text at the initial stage of analysis, but they are constructing syntactic structures within the text in order to get to a semantic level of processing.

Whilst it is necessary to acknowledge the ever increasing output of research work and results from the area of computational linguistics generally, it should not be forgotten that my work is directly concerned with identifying features (or properties) of text which imply the organisation of information therein. What I have called conceptual information is that which most research work in computational linguistics is concerned with, for those projects of thematic and stylistic analysis. The projects concerned with meaning representations and the like are probably more strongly related to my work. The former and latter categories may be grouped under literary and linguistic research respectively, but because of the common base of structural analysis, particularly when using a computer, the two are by no means mutually exclusive and from the literature often seem to cross boundaries. The key to how work in literary and linguistic computing relates to this thesis, is in the term structural analysis.

I have attempted in this research to limit my work to the formulation of a grammar consisting of semantic labels which can be used to classify statements in text which tell us something of the organisation of authors' messages. I have refrained from attempting computer analysis of text using the grammar because I feel that there is a complete topic there on its own. More will be said of future work using the grammar in Chapter Six. Structural analysis can only be carried out after the structure has been defined and a grammar or other analytical tool has been developed. Much of my research, therefore, must be in the area of linguistic structures, with a view to analytical implementation and techniques for carrying out same. To review and give an evaluative report of all the current research in this area is, I think, warranted. There are however some major problem areas and research into those which I should mention in addition to what has already been discussed. As HOCKEY (1978) records though, there is still much dispute over methods and techniques for computer use in literary and linguistic research. In this report of a colloquium on textual criticism using computers, Hockey summarises several papers which outline methods for literary criticism by computer and implies that a major problem is editing in semantic markers for further text processing - this occurs after the identification of relevant 'chunks' of text for authorship comparisons and the like, which in itself is a many-faceted problem.

Predicting sentences or concepts from knowledge already acquired during

processing is an area which may be helpful to me when trying to classify statements of meta-information in relation to the conceptual information they are associated with. LANGFORD and HOLMES (1979) have studied the syntactic presupposition of subjects in experiments with sentence comprehension. They set up target sentences and gave subjects 'base' sentences from which to attempt target sentence comprehension. They concluded that at the syntactic level they were working at, this methodology was appropriate and may have further use at a semantic level. Whilst at the syntactic level though, ORNAN (1978) discussed the problems of different word forms and their generation by computer without using a dictionary. Methods for this kind of process would certainly be of use to me when analysing a highly significant meta-informational word such as describe, which may have a number of forms depending on its tense in context. It could be describes, described, describing, description or descriptive at least.

Another dispute in linguistics which I can relate to and comment on for this project, is that of the relationship of semantic structures to syntactic structures so far as their linearity is concerned. My grammar defines a three phase linear structure for empirical argument in science text. Any structures produced therefore, are by nature linear. I will show in Chapter Four that science text is not as linear as might be imagined, but can and is linearised when summarised with and without my grammar. BARTSCH and VENNEMANN (1973) say that an unwarranted assumption from 'generative semantics' (also from Chomsky's transformational grammar), is that logical forms which represent sentences are linearly ordered. This may be so if one is only considering the relationship of word order to its logical form representation, but not so if we are determining the existence of some structural representation of arguments, as I am attempting to do. The argument structure is related to the 'message' being communicated by the author and is thus yet another level deeper in a semantic sense. If we were only discussing the formal nature of argument structures I would have to agree with McCAWLEY (1971) who says that an exact equivalence exists between the nature of deep structures and surface structures. Both can be represented by trees "whose non-terminal nodes are labeled by symbols interpretable as syntactic categories" (p.221). CHOMSKY (1965) proposed this to some extent with his deep structures theory. This is quite important for any future implementation of my grammar because in order to process the 'chunks' of text which

have been assigned category descriptors, I have to treat the resulting structure in a syntactic way if I want to for instance, generate a summary of whole text by merely concatenating the 'chunks' of text thus classified with the grammar. At this point I must treat the text syntactically, but the order of the structure has already been predetermined semantically by the grammar itself. To this end I reject BARTSCH and VENNEMANN's comment that the relationship of linear forms between surface and deep structures is unwarranted; it is a matter of fact that ordered forms are necessary to produce coherent text, even if the semantic relationships between aspects of say an author's argument (message) are non-linear. By removing redundant text in the classification process the text can be linearised anyway - see results of experiments in Chapter Four.

Work with semantic structures is an increasingly popular topic in linguistics, psychology, computers and information science. As I have attempted to show, perhaps the major conceptual problem for computational analysis of text, is how to recognise discrete semantic categories in 'chunks' of text and their relationships to one another in for instance, arguments or messages. WERLICH (1976) gives a comprehensive list of semantic properties in English text and is a useful reference if one is deciding on which aspects of text should be extracted to create a particular representation of same. I have not given a bibliography of works in this area, but the references listed in this thesis should give enough pointers to a sufficiently representative corpus of literature in the area of computational linguistics to show the relevance of that field to my problem and Information Science.

Quite simply, there is one major linguistic problem which faces any computer implementation of my grammar. That is, the recognition of 'chunks' of text (be they phrases or statements) which imply classification by one or another of the grammar elements. A grammar which defined a segmentation of text into say, ⟨First Part⟩ ⟨Second Part⟩ ⟨Third Part⟩ ⟨Fourth Part⟩, could be implemented very crudely by searching for words or phrases which said,

"First..." or "To begin with..."

"Second..." or "Next..."

"Third..." or "The second point is followed by..."

"Fourth..." or "lastly..."

To some extent this is a valid approach because words like "First" are appropriate to the structure being generated. My grammar attempts to go beyond this definition of a semantic structure to show how empirical arguments are organised in science text and future implementations will, I hope, use more sophisticated categorisation procedures such as statement prediction mentioned by LANGFORD and HOLMES (1979) above. More is said of this in Chapters Three and Six.

NIDA (1975) advises one to avoid asking, "What does the term mean?" (p.169), when determining meaning of lexical units. He suggests it is better to ask "What is it like?", "How is it used?", "When do you say this word?" His work entitled Componential Analysis of Meaning provides many useful criteria like this for analysing text to produce a variety of semantic structures. Semantic domains and what he calls 'extralinguistic entities' in text can be categorized for particular representations of the text. In fact, he gives a very comprehensive list of examples of semantic domains in a generic structure which spans nine pages of his book. The existence of basic works like this are essential to the study of semantics for Information Science as well as for linguistics and other fields. One of the most relevant implications from Nida's book for my work is his reference to ordered relations between components of meaning. For instance,"repentence shares with remorse the component of contrition..." (p.34), which is one of his own examples. Establishing these kinds of relationships between meta-informational words or phrases in text would help me in deciding whether to include a given 'chunk' of text in a structure which summarises it, for example. This kind of work will also form later investigations beyond actually establishing the form of the argument structure, which is mostly what this research has attempted to do.

2.5 SUMMARY

This chapter has been an attempt to point to some of the major work in semantic information processing in the context of my work and the development of the whole field. Some difficulties of processing natural language text have been discussed, such as ambiguity and anaphora in text, and the emphasis has been on automatic document analysis, especially for MT. The reasons for this are that MT through practical experience, has given rise to many of the fundamental problems we are now trying to deal with. The overall 'message' of this chapter has been that over the past twenty years some progress has been made, but very little in terms of reliable operational systems for any form of automatic natural language processing. Although my work is primarily concerned with developing a set of semantic labels which can be used to represent the organisational structure of author's arguments in empirical science text, the work described in this chapter is the foundation from which my nomenclature and methodology comes.

CHAPTER 3

A META-INFORMATION STRUCTURE FOR AUTHORS' ARGUMENTS

3.1 OVERVIEW OF THIS CHAPTER

The last chapter dealt with the general development and problems of semantic information processing. I now want to return to some of the issues raised in Chapter 1 and in particular to the assumption underlying this thesis which is described in section 1.3. In this section I outlined my assumption that there was an ideal in the minds of both writers and readers of science text, about the existence of a conventional format for the presentation of empirical arguments. I gave some intuitive and observational evidence for the existence of this belief. Although section and paragraph headings in a large number of text which I analysed seemed to follow the notion of this conventional format, I said that we now needed to examine individual statements to establish their distribution patterns. To do this requires that a set of labels be constructed with which we can classify individual statements. This set of labels needs to be representative of the consensus view of the conventional format.

In this chapter I will outline the set of semantic labels I have developed to classify individual statements and the rules used to carry out the analyses. First however, I shall endeavour to give my view of the term meta-information in the context of previous definitions and concepts of information itself. I hope to show why the organisations which I produce from text analyses are best thought of as meta-informational entities.

3.2 THE CONCEPT OF META-INFORMATION

For my purposes, arguments in empirical text have two distinct properties. On the one hand they convey conceptual information to readers in the form of subject knowledge, ideas and concepts. On the other hand, the argument itself is organised or presented in some way which is itself indicative of the communication process which takes place when a reader reads the text. This organisational facet of the argument is inferred semantically by the combination of the writers own words on the paper. Phrases such as 'This paper sets out to ...', indicates the intention or aim of the argument, for instance. If the conceptual information is the 'message' of the argument or the 'actual information', then I suggest that the semantically inferred organisational information can best be thought of as meta-information, for it describes the conceptual information. A definitive theory of meta-information by SHREIDER (1974) shows how we can separate actual information from information which describes the actual information. He says, "... as soon as we become interested in the context level, (the use of the information contained in the text), we are immediately faced with the problem of the need to know the connection between the information and the text. This knowledge, (information about the method of coding the information in text), is what we shall call meta-information." (p.3). My interpretation of what he is saying here, is that we must be able to identify elements in text which tell us about the conceptual information therein. I believe that in practice, SHREIDER's theory comes down to semantic inference in text and if we can label that then we have a meta-language in effect. My set of labels for classifying statement types with empirical argument fulfils, I feel, this criteria.

Several other contemporary researchers have expressed opinions regarding the nature and being of information. See DEBONS (1974) for an overview of other concepts of information for Information Science. For my own part I prefer BELKIN'S (1977) interpretation when he endeavours to establish an integrated concept of information for the study and solution of problems within Information Science. He proposes a theory which is based on the belief that an invariant structure exists which is associated with all text. This structure, he says, can be thought of as the information pertaining to that text. Belkin's use of 'associative structures' (pp:129-161) to show how this information is

directly relatable to the conceptual structure of the originator of the text is most useful to my work, because it suggests, (if only by indirect inference), that the conceptual structure of a recipient may also be related to that text. My interpretation of the invariance of Belkin's information structure, is that a number of rules must exist for the generation and interpretation of the structure. In this way the rules of the grammar which represents the notional conventional format which I have mentioned, would be applied to the writing or reading of text. I was interested to see whether or not the conceptual structure of readers of science text could be directly related to the information structure in text.

A final reference to contemporary concepts of information for problems in Information Science comes from BELKIN and ROBERTSON (1976). In this paper, information was defined as being, "... the structure of any text which is capable of changing the image-structure of a recipient" (pp: 199). This has two aspects of interest for me. First, the 'image' or conceptual structure of the recipient is mentioned. This is of prime significance bearing in mind my assumptions concerning the transfer of information from text to the recipient. That is, if the structure is the information, then it is the structure (or output) from parsing the text with the grammar, that is being transferred. The second interesting aspect of the Belkin-Robertson definition is that it infers an ability on the part of the information from text to actually change the state-of-knowledge of the recipient. Although this may seem a common sense assumption, it is worthy of note in light of the problems of trying to define what is transferred as information from text and how the recipient's state-of-knowledge is changed by that information.

To delve deeper into the nature and being of information is not warranted. All that is required here is to illustrate the assumptions and definitions that are being used as a theoretical base for this research. The fact that they are being used is an indication that an integrated approach to problem-solving within Information Science research is becoming a reality.

3.3 A SET OF SEMANTIC LABELS AS META-INFORMATION ELEMENTS

The term grammar as I use it here and as it has been used by linguistics and other disciplines such as psychology, has evolved from the natural language use to mean rules for reading and writing text and labels to describe parts of speech; in syntax the labels noun, verb, adverb and so on, and in semantics the labels event, result, and so on. The descriptive semantic labels of my grammar are given below and are presented in the form of a production chain. This means that any label on the left-hand side of the production symbol (: : =) can be substituted for by the labels on the right hand side. The vertical bar (/) means that either one or the other or both of the labels on each side of the vertical bar can be used. The effect of this description (Backus-Naur Form (BNF), referred to in Chapter 2) is to give a clear meaning to how labels can be used. In fact, this production chain means that every argument should have three phases in the order stated (linearly) but within each phase any one or all of the statements described can appear. I outlined this production cahin and a partial-parsing algorithm for its limited implementation with natural language text in SALLIS (1978b). In this paper I pointed out the properties and elements of the grammar and the constraints placed upon it by the inherent variability and ambiguities in natural language text. I also discussed two possible approaches to algorithmic text analysis and these are outlined in the following section of this chapter.

The following grammar shows the elements of empirical argument as may appear in a variety of combinations within science text. This treatment of empirical arguments to produce an information structure which reflects the organisation of the arguments in science text, implies that one overall or macro structure exists for each text.

A development of the structure could incorporate infra or micro structures within the macro structure. Without any further evidence at this stage, I assume that the micro structures would consist of similar elements to the macro structure, which I have described below as having three phases. That is,

⟨emprical argument⟩::= ⟨phase one⟩ ⟨phase two⟩ ⟨phase three⟩

The macro structure would still exist to represent the overall argument or 'message' of the text, but a number of micro structures could exist within it. We would therefore, define text as having:

⟨empirical argument⟩::= ⟨micro structure⟩* /
⟨phase one⟩ ⟨phase two⟩ ⟨phase three⟩

⟨micro structure⟩ ::= ⟨phase one⟩ ⟨phase two⟩ ⟨phase three⟩

where the asterisk (*) means repetition. We could define this in another way by stating that,

⟨empricial argument⟩ ::= ⟨macro structure⟩/⟨microstructure⟩*

and then further define both macro and micro structures as having⟨phase one⟩ ⟨phase two⟩⟨phase three⟩elements. Either description would be equally valid.

The experimental results discussed in Chapter Four do suggest that there is more than one structure for argument organisation in the texts analysed, which would seem to support my assumption that this is generally true. More is said of this in Chapter Four but it should be noted that the main purpose of the experiments and indeed the thesis, is to demonstrate the existence of an overall meta-informational structure in text which can be represented by use of the semantic labels in the following grammar.

FIGURE 7

PRODUCTION CHAIN OF SEMANTIC ELEMENTS

⟨Text⟩ : := ⟨Emprical argument⟩

⟨Empirical Argument⟩ : := ⟨Macro structure⟩ / ⟨Micro structure⟩*

⟨Macro structure⟩ : := ⟨Micro structure⟩

⟨Micro structure⟩ : := ⟨Phase One⟩ ⟨Phase Two⟩ ⟨Phase Three⟩

⟨Phase One⟩ : := ⟨Phase One Statement⟩ /
⟨Phase One⟩ ⟨Phase One Statement⟩

⟨Phase One Statement⟩ : := ⟨Introduction⟩ ⟨aim⟩ ⟨hypothesis⟩
⟨observation⟩ / ⟨assumption⟩

⟨Phase Two⟩ : := ⟨Phase Two Statement⟩ /
⟨Phase Two⟩ ⟨Phase Two Statement⟩

⟨Phase Two Statement⟩ : := ⟨data⟩ / ⟨method⟩ / ⟨evidence⟩ / ⟨citation⟩ /
⟨result⟩ / ⟨evaluation⟩

⟨Phase Three⟩ : := ⟨Phase Three Statement⟩ /
⟨Phase Three⟩ ⟨Phase Three Statement⟩

⟨Phase Three Statement⟩ : := ⟨Conclusions⟩

As mentioned earlier, this is a full production chain of the grammar which assumes that there are micro structures within the macro structure and that the micro structures are identical with the macro structures. Only a sub-set of this grammar was given to subjects to use in the experiments described in Chapter Four. That is, the existence of micro structures was not suggested to the subjects for any analysis in the experiments. The labels within phases of the argument structure were also given to subjects in a less formal way, as can be seen in Appendix C and D.

The grammar is by definition a set of rules for how it should apply to a given text. In this case a linear structure, ⟨Phase One⟩ followed by ⟨Phase Two⟩ followed by ⟨Phase Three⟩, is placed like a template on top of a text which may be linear or non-linear. WILKS (1976) might call this template a frame. Whether or not the text is linear, some re-ordering of 'chunks' of it will probably be necessary to conform to the grammar. Still more rules are required so that this re-ordering can be carried out though, because the grammar does not define what a 'chunk' of text is - I have said it could be a word, a phrase, a sentence or a statement. More precise definition of 'chunks' is necessary for a computer implementation of the grammar, but for purposes of this project where humans are conducting the text analyses, guidelines have been produced. LORD (1974) said, "If any linguist sets out to provide a complete descriptive grammar of a language, he is doomed to disappointment. The nearer he seems to be getting to his goal, the more numerous are the features that refuse to comply with his rules or fit into his system. He is forced to create rules and still more rules, until his grammar reaches the point of becoming hopelessly unweilding and uneconomical. The end result would seem like setting up a pile-driver to crack a nut". (pp: 197-198). I am not at all sure that one complete grammar is sufficient for the analysis of text which I am eventually hoping to conduct by computer. That is not to deny that it is possible and perhaps appropriate for grammars of say the English language, but after all the structures which I am concerned with are more descriptive of one semantic 'view' of text, not a whole language. In any case, the grammar I propose here does not endeavour to do anything more than show the descriptive terminals of a structure which could represent arguments in empirical text. I think it is important to realize that, whilst considering the potential of the grammar, in conjunction with some implementation rules, for whole text analysis.

This set of labels reflects the three-phase notional format arrived at by consensus, intuitive and observational evidence. Its validity or appropriateness to my kind of analysis becomes apparent in the next chapter where I discuss results of text analysis by a large number of individuals.

Although the consensus view of the conventional format assumes a three-phase progression for empirical argument, there has been no experimental evidence from my work to suggest any precedence for the order in which the labels should appear in the production chain. In order to demonstrate the kind of statement that could be classified by the individual elements of the production chain in Figure 7 above, I have listed some examples below, (Figure 8).

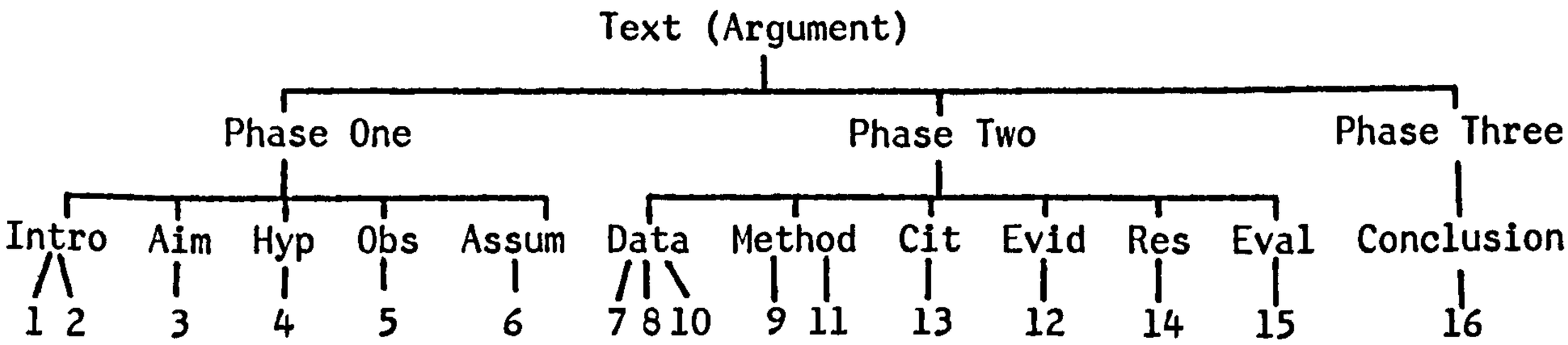
The first aspect of this example to take note of is the splitting up of sentences into statements. Take the first two statements for example. Both are from the same sentence and joined by the conjunction (and). In this case both statements have been classified with the same label but that is not always necessarily so. The use of the same label twice demonstrates a further application of the set of labels. They may be used any number of times in one analysis. As shown in statements 9, 10, and 11, they may be used in any order within phases too. The results of text analysis in the next chapter show how seldom the distribution of labels actually gets to the example given in Figure 8. The example demonstrates one more important factor when considering the construction of rules for classifying statements with the set of labels as discussed in the next section. That is, very few of the statements contain words which are equivalent to the labels themselves. For instance the word 'hypothesis' actually occurs in the statement before the one classified with the label ⟨HYPOTHESIS⟩. Therefore straight word matching is considerably unreliable in this kind of analysis. From this example, I hope it is clear how I arrived at the two kinds of information I mentioned earlier. The conceptual information in the text of Figure 8 relates to the argument about fish tank evaporation. The meta-information is the semantically inferred organisation of that argument. No matter how these statements are arranged, the same labels would be used to classify them. What the conventional format assumption says is that the 'ideal' organisation of an argument is something like the one shown in the Figure 8 example.

FIGURE 8

EXAMPLE STATEMENTS AND THEIR CLASSIFICATION

LABEL	SEQ. NUMBER	STATEMENT
Introduction	1	This paper relates to my research in fish tank evaporation (and)
Introduction	2	I intend to discuss my progress to date.
Aim	3	I set out to test the following hypothesis:
Hypothesis	4	that water evaporates at temperatures below 20°C.
Observation	5	I proposed this after noticing how the water level in my fish tank dropped even during very cold weather.
Assumption	6	I assumed that the drop in water level was not due to an overt thirst by my fish.
Data	7	My tank holds fifty gallons of water.
Data	8	Most days I was topping the tank up with about two gallons.
Method	9	On Monday night I marked the tank after topping it up;
Data	10	I know it now had fifty gallons in it.
Method	11	Every night for two weeks whilst the average temperature was 20°C, I measured the amount of water it took to top the tank up.
Evidence	12	My neighbour, who is a Magistrate, recorded my activities on data sheets which were later analysed.
Citation	13	This is the classical method of data collection for fish tank experiments proposed by WHOKNOWS (1973).
Result	14	The average amount of water taken to top the tank up was two gallons.
Evaluation	15	This result is consistent with my previous experiments.
Conclusion	16	The water loss can only be accounted for by evaporation.

Perhaps a useful way of presenting the argument structure of the text in Figure 8 is by using a tree-diagram. The result of this would be:



The numbers are the sequential-numbers of statements in the text which contain conceptual information. The labels represent the meta-information.

3.4 RULES FOR TEXT ANALYSIS

In Chapter 2 I demonstrated the problems of ambiguity in natural language and discussed the case of anaphoric reference as a 'stumbling block' to computational text analysis. The real problem for computer analysis of natural language is that whereas computer programs require precise and logical rules in order to code algorithms to carry out such parsing, natural language itself is so variable that rules for general-purpose use are virtually impossible to create. SCHANK (1977) in his work on rules and topics in conversion, maintains that although rules must be invariant to be rules, they should also be flexible within the topics which they relate to. Some topics of conversation allow us to make sparing use of rules for ambiguity because we have a central issue to discuss which can be referenced by precise terms. These terms might otherwise be ambiguous in general conversation. LEVIN and MOORE (1977) also made similar assumptions in their study of dialogue games: meta-communication structure for natural language interaction. They discuss categories of semantic inference which can transcend the usual or generally accepted meanings of terms and phrases. These notions are major advances in the field of natural language processing because they help us to cope better with the problems of ambiguity and anaphora which always present exceptions to rules.

An interesting view of 'probablistic' or 'utitlity theoretic' indexing has recently been published by COOPER and MARON (1978). Their paper outlines their endeavours to establish a rational criterion for human indexers to decide on which indexing terms or descriptors they should assign to a 'unit' of stored information for purposes of later retrieval. A 'unit' can, I assume, be any element of stored information from 1 to n characters long. The interesting aspect of their paper for my topic is their discussion and production of explicit decision rules for both kinds of indexing. This unified theory of indexing is expressed from a common conceptual and mathematical foundation. Using a normative approach Cooper and Maron say what indexers should do, not what they are doing and how these contemporary practices can be interpreted then perhaps modified. They construct a probability model from data relating to previously used terms and their frequency of use, then assign weightings to each. Others have done this previously. For example, see SPARCK JONES (1978). Later, Cooper and Maron apply boolean operators to the terms and manipulate them for the 'best fit' when retrieving their units of information. The really important aspect of this work is that these individuals are more concerned with the rules and descriptive elements than with the retrieval strategy. This former area is the one which I feel requires most work at present - the formulation of rules and meta-linguistic entities to describe aspects of text.

Instead of algorithmic rules I have tried to produce some generally accepted and useful 'guidelines' for text analysis using my grammar. Figure 9 which follows shows the 'guidelines' given to subjects in the experiments described in Chapter Four. These rules were produced after a great deal of trial and error and the use of a tape recorder by subjects when attempting to classify statements in text using my grammar.

FIGURE 9

'RULES' OR GUIDELINES FOR TEXT ANALYSIS

1. Only use one label for each statement.

2. Labels may be used more than once in any order.

3. As a general rule, treat sentences as statements but obviously some sentences contain more than one statement. Make your own decision as to what a statement is, but to make your choice obvious, separate statements by putting a sequential number in front of any statement you discover.

4. Where ambiguity exists in your choice of label, put an asterisk followed by the labels you consider to be applicable.

5. Only classify assertive statements or propositions NOT questions for instance.

6. Try to identify statement-types by indicative words or phrases. That is, the phrase 'the aim of this paper is to ...', would indicate the use of the label ⟨aim⟩. If no direct relationship exists between words in the statement with labels in the grammar, use thesaural-type relationships to determine the classification. For example, the term 'outset' is related to ⟨introduction⟩. In all cases use your intuition to determine the semantic inference of a statement.

7. Where anaphora or other reference from one statement to another occurs, try to classify the statement in question as a separate entity, from the other statements around it by substituting say, pronouns with the noun being referred to.

Although these are only a guide to individuals attempting text analysis, I received a favourable reaction to them from candidates in the experiments. Overall, individuals have to use their intuition to classify statements in the way I suggest. In many cases there are no syntactic indicators such as equivalent words - semantic inference is the key to statement-type classification.

A reader's knowledge-store has the pragmatic advantage that a computer program does not and the many facets of it can be employed simultaneously when text is being read and interpreted. As shown in Chapter 2 some work in Artificial Intelligence is making advances in

this area of text processing, but fool-proof systems are as yet a long way off. The human mind can carry out several processes at the same time and do this heuristically. Heuristics enable the analyser to back-track and make new decisions based on current experience. The infinite variation of written natural language makes algorithm design to cope with this kind of phenomenon virtually impossible. I have already mentioned some of the types of statements which can be classified within each of the phases of argument. Statement-types can be recognised either deterministically by the use of function words (words which denote semantic associations with one or another statement-type in text), or probabilistically by establishing the position of statements relative to one another in the text; for instance, a statement of conclusion usually coming at the end of an argument and thus the end of a text. It may also be possible to use some probability estimations of statement-type distribution based on the frequency of co-occurring words in text. Work by BEKTAEV (1977) and other Soviets using this approach for micro-glossary construction, often using syntactic markers in text like punctuation as statement determinators, may have possibilities for semantic categorisation beyond their present essentially lexicographical applications.

Subject terms can be recognised by matching words in statements with the thesaural part of a reader's knowledge-store. That is, the subject state-of-knowledge can be represented by a thesaurus, or at least the vocabulary of that subject can be so represented. Therefore, the notion of a thesaurus is a useful one for discussing this facet of a reader's knowledge-store. No actual subject facet analysis is being attempted here. The only purpose in discussing subject-term recognition is in order to acknowledge that readers obviously do carry out some sort of parsing like this during the interpretation of a text. The conceptual information in an author's argument must obviously contain a great deal of subject vocabulary, but we are only interested in representing the organisation of this argument, or in other words the overall organisation of the information in text as it relates to the author's argument which is being communicated.

The idea of a thesaurus of terms which can be matched say by words from text, need not only be subject-related. Take, for example, the recognition of words which denote the concept of result from the grammar. Rogets Thesaurus gives the following list of terms as being synonymous with the word result.

FIGURE 10

THESAURAL SYNONYMS FOR THE TERM 'RESULT'

sequel	consequence	effect	conclusion
end	aftermath	legacy	postlude
epilogue	postscript	finish	completion
colophon	coda	termination	derivation
remainder	upshot	outcome	issue
eventuality	product	output	end-product

If we were using the above as a stop list of terms to the nature of statements containing one of them as being a statement of result, we have one immediate problem. That is any of these terms might occur as incidental words in another kind of statement. On encountering such a situation using this stop-list method, statements are likely to be incorrectly classified. Another problem arises in the compilation of a list such as this in that single words rarely determine the type of a statement in or out of context. The inferences are more semantic than the mere meaning of one word. It is the combination of the words and their relationship to one another which implies the nature of statements in text. In any case, even after the massive compilation task, the arbitrary nature of lists of terms like the one in Figure 10 above prohibits reliable text analysis. The use of lists and dictionaries of words remains an integral part of many text processing systems, particularly with machine translation where term equivalences in various languages are sought.

3.5 SUMMARY

This chapter has had two specific goals. The first was to give an explanation of my use of the term meta-information and reasons for using it in the context that I have. The second goal was to outline and describe the full production chain of semantic labels which I have developed to represent the consensus view of a conventional format for empirical argument presentation. A short discussion of the analytical

rules which were given to candidates with the set of labels when they attempted to classify individual statements in text, has also been given. Overall, it is important to realise that no reliable set of algorithmic rules has been successfully constructed. The results from experiments given in the next chapter are the result of human analysis and have been interpreted as such. In short, this chapter has attempted to give the ingredients of the structures which are produced from text analysis.

CHAPTER 4

EXPERIMENTS IN TEXT ANALYSIS

4.1 OVERVIEW OF THIS CHAPTER

This chapter outlines the results of the experiments which were conducted in this project using my grammar as discussed in the previous chapter. There are five experiments in all. The first two are the result of a 'pilot study' which I carried out before undertaking a larger scale approach to the experimentation.

In the first experiment, (section 4.3), I used the grammar to classify individual statements within the introductory sections of a number of text. This exercise was aimed at establishing whether or not all the statements within a section of text with a particular heading were in fact of the type suggested by the heading. This experiment is referred to as 'Part I' of the Pilot Study.

Part II of the Pilot Study, (section 4.4), was an experiment where three subject specialists were given a text in their field of teaching and research which had no section headings. They were asked to identify individual statements in the sample text, then label each statement with one of the grammar elements. The purpose of this experiment was three-fold. First I wanted to see whether my grammar could be used by others as I had used it. Second, whether there was any agreement by all three participants as to which portions or chunks of text were statements. Third, whether there was any agreement between the three participants as to which grammar elements should be used to label individual statements.

The results from the Pilot Study were encouraging enough for me to embark on further experimentation which is described in

sections 4.5, 4.6 and 4.7. In section 4.5 I describe an experiment where a sample text is classified by a group of 21 scientists in the same manner as the Part II Pilot Study was conducted. The results of this experiment are compared in section 4.6 with an experiment using the same sample text and methodology with a group of 21 non-scientists.

Having given these results, one final experiment was conducted using the sample text from the above mentioned two experiments. In section 4.7 I describe how two groups of scientists were asked to produce a summary of the author's argument in the sample text. One group of five subjects had prior knowledge of my grammar and the conventional format thesis, but the other group of five did not. A comparison between the structures produced from the summaries of both groups of participants in the experiment was made and appears in section 4.7. An overall interpretation of the results is given in section 4.8. The chapter finishes with a summary of the contents of the experiments and their results.

4.2 EXPERIMENTAL METHOD

The purpose of this experimentation is two fold. First, there is the overall question of what kind of information structures can be produced from text using my set of semantic descriptors as organisational elements to reflect the presentation of authors' empirical arguments. Second, there is the applicability or 'appropriateness' of my set of descriptors for this task. In the overall scheme of this thesis though, both of these questions only represent one facet of the topic being investigated. I am, after all, attempting to demonstrate a theory of information transfer as well as information organisation. Having classified statements in text using my grammar, to show the kind of information structure which can represent the organisational facet of the problem, the transfer of that information from text to readers becomes my next concern. To cater for this aspect of the problem I used readers to produce summaries of a sample text in an endeavour to determine how they synthesised and interpreted structures which I knew existed in the text given to them. In short, I wanted to see whether the readers of a text which was judged to

have a structure of one kind representing the authors' argument would produce summaries with similar or disparate structures. Therefore, although most of the experimentation described here is concerned with classifying statements in text using my set of semantic descriptors, it is also aimed toward discovering something about the process of information transfer from writer to reader through the medium of natural language science text.

Before each experiment there is a preamble about its nature and operation. At the end of each are some conclusions about the results presented. The last section (4.8) endeavours to interpret the results given with a view to providing data for a model of the information transfer process which is at the root of this work and which is outlined in the next chapter.

When evaluating the experiments here, particularly the first one in section 4.3, it should be remembered that part of my initial data for formulating an assumption about the existence of a conventional format, was that out of 150 science journal texts examined, 131 had section headings which followed the general format of the grammar. This was the 'starting block' for attempting to ascertain the distribution of statement types in text to see whether the individual statements actually followed the semantic inference of the section headings, let alone the grammar.

I have differentiated between 'scientists' and 'non-scientists' in the experiments and should define the difference between the two groups here. The 'scientists' referred to are all graduates in pure or applied science and the 'non-scientists' are all professional indexers or abstractors in scientific and technological literature with experience ranging from five to fifteen years. Thus, the first group have subject knowledge and are familiar with disciplines which use empirical argument as an integral part of their methodology. The second group have knowledge and practical experience with structuring lengthy text into summaries which reflect the essential aspects of authors' arguments.

Other aspects of the experimental method are related to the individual exercises which are described below.

4.3 THE PILOT STUDY - PART I

4.3.1 Comparing statement-types with section-headings in science text

The purpose of this experiment was to take a particular section heading in a number of science journal articles and attempt to classify individual statements within the chosen section of each text to see whether the statements were in fact of the same type as the section heading. The heading chosen here was 'Introduction' and by virtue of the rules of my grammar, any statement which can be classified using one of the semantic labels in ⟨Phase One⟩ of the grammar, (which is where ⟨Introduction⟩ resides), is deemed to be of an 'introductory type'.

The text used here were all chosen from the sample of 150 previously mentioned. There were 20 texts chosen with introductory type headings. I decided to produce self-documenting analyses by plotting my results on x/y graphs. As each text was read I classified individual statements as I recognised them, using whatever label seemed appropriate from anywhere in the grammar. The graph contained sequential numbers on the horizontal axis for the individual statements as they occur in text and on the vertical axis the grammar labels were dispersed within their phases. Having classified each statement within the 'introductory' section of the text, the co-ordinates which I had appended to the graph were joined by a continuous line from the first to last statement number. If the co-ordinates joined by this line kept within the 'introductory' phase of the grammar, I considered the result to be of a linear nature and within the conventional format ideal. If the co-ordinates, (that is the statement-type distribution), left the confines of the 'introductory' phase, I considered this representation to be non-linear and non-correlative with the conventional format ideal. This explanation is perhaps made clearer by the diagram (Figure 11) shown with the results.

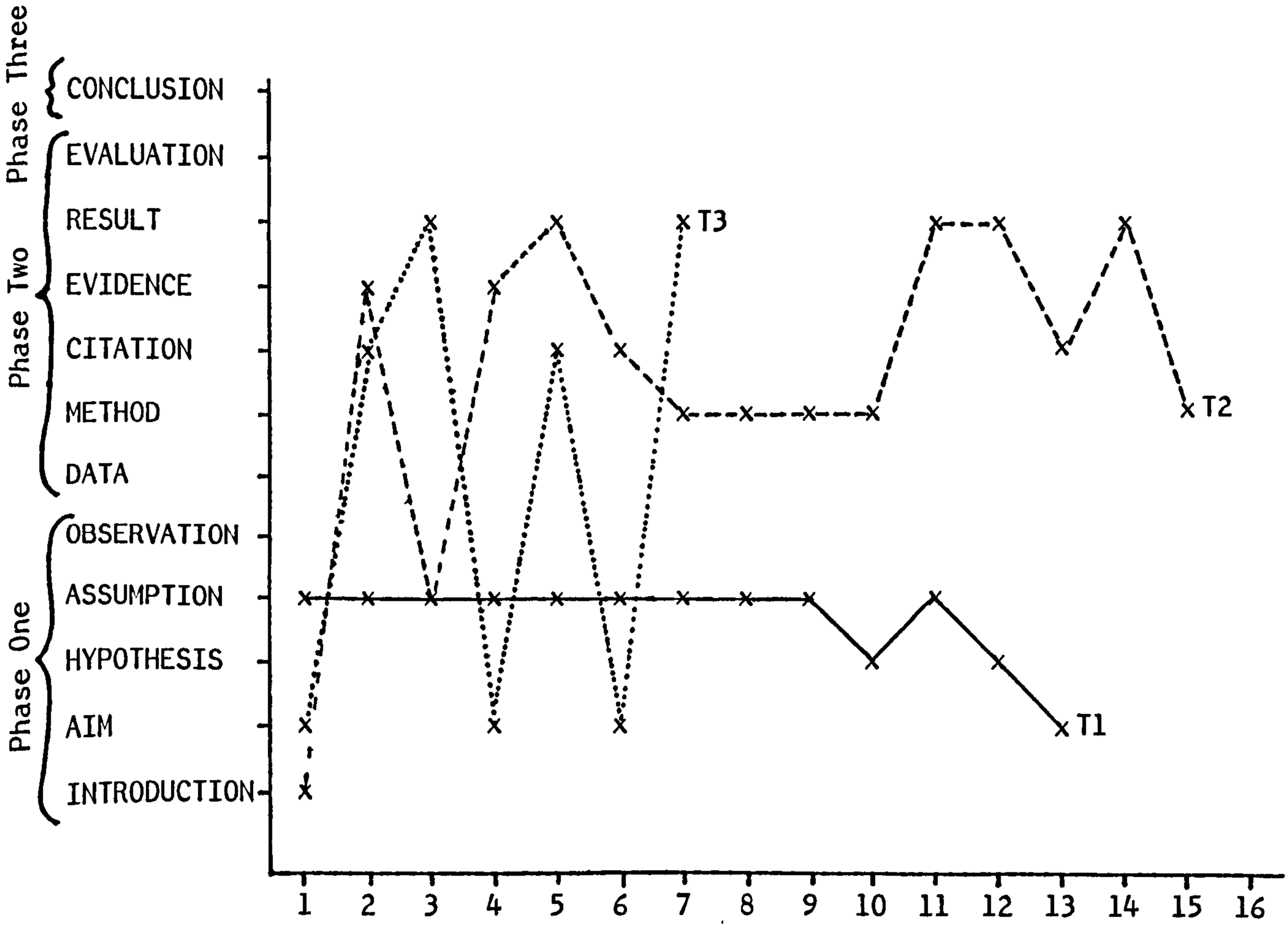
4.3.2 Results of the experiment

The overall result of my analysis was that 17 out of the 20 texts displayed statement type distributions which were non-linear and

therefore not in line with the conventional format ideal. That is to say that 17 of the 20 texts did not have all the statements within their 'introductory' section classified by the label ⟨introduction⟩ or any of the ⟨Phase One⟩ homous labels of ⟨hypothesis⟩, ⟨aim⟩, ⟨observation⟩ or ⟨assumption⟩. I shall give my interpretation of this result shortly but in the meantime I have shown in Figure 11 the distribution of statement-types for three texts which formed a part of my sample. T1, T2 and T3, refer to three different texts.

FIGURE 11

GRAPH SHOWING COMPARATIVE STATEMENT-TYPE DISTRIBUTIONS



As can be seen here, T1 stayed within ⟨Phase One⟩ of the grammar and so is judged to have a linear representation in line with the conventional format ideal. The texts T2 and T3 obviously do not

have linear representations. The number of statements in each text is irrelevant for comparative purposes at this stage because it is the pattern of each analysis and the form of individual structures which we are primarily concerned with, not the number of statements which individuals recognised. The remaining seventeen graphs showing results from this experiment can be seen with a short explanation of each in Appendix B. One of the most obvious factors in common of those three texts which were linear, is that they are all short and have few statements, which suggests that lengthy text is difficult to keep within the scope of its section-headings.

Interpreting the result from this experiment is perhaps as subjective as the analysis which formed the core of it. An immediate reaction might be that the grammar labels should be examined to see whether they are the most appropriate for the task. I think that a close examination of the graphs is more fruitful though because we can see for instance, that statement three (3) in text T3 has been classified as a statement of result and unless my intuition is at fault that could hardly ever be fitted in any of the ⟨Phase One⟩ categories. Of course my (or anyone's), intuition or judgement may be at fault, but after all we are dealing here with a human process and therefore we cannot place any absolute controls on this kind of activity. If however, we can get sufficient agreement on the classification of individual statements using particular labels, we might eventually be able to develop algorithms to conduct this kind of analysis; at which stage a computer could do it for us.

4.3.3 Conclusions from Part I of the Pilot Study

My overall interpretation of this result is that although writers, (and in fact journal editors), have an intuitive ideal of a conventional format for presenting empirical argument in text and they try to enforce it by using section headings which follow this format, individual statements are not distributed in the same linear fashion. This could be for a variety of reasons of style, expression and the inherent ambiguity of written natural language. For my purposes the interpretation must stop there. My concern is purely with the organisation of the argument and in identifying the meta-informational structure which reflects that organisation. Taken individually, each

of the results in the graph in Figure 11 is a representation of the meta-informational structure of the text, reflecting the organisation of the author's argument within it. Before moving on to subjecting one text to a large group of individuals for analysis, I first wanted a small group to analyse a text without section headings to examine their results. Thus, I embarked on Part II of this Pilot Study which is described in the next section.

4.4 THE PILOT STUDY - PART II

4.4.1 Classifying statements in sample text with a small number of subjects

This part of the project marks the beginning of the real experimental work. The aim of this exercise was to see whether other individuals (three subject specialists in this case) could use my grammar to classify statements in a sample text and to see just what kind of structures were produced.

The text was chosen for its short length (1079 words) and absence of section-headings. A copy of the sample text appears in Appendix C, with a copy of the instructions given to the three subjects.

Each subject was asked to identify individual statements in the text by putting a sequential number in front of each new statement found. For each statement they were asked to choose a code-number from the grammar elements and record it as the classification code for that statement. The codes for elements of the grammar are given in Figure 12.

FIGURE 12

CLASSIFICATION CODES FOR GRAMMAR ELEMENTS

CODE	GRAMMAR ELEMENTS
1A	Introduction
1B	Aim
1C	Hypothesis
1D	Assumption
1E	Observation
2A	Method
2B	Evidence (Data)
2C	Citation
2D	Result
2E	Evaluation
3A	Conclusion

4.4.2 Analysing the Results

Having received the results from the three subjects, I first numbered each word sequentially in the text they had used. This enabled me to construct a list of word-spans or co-ordinates for each subject, by noting the numbers of the first and last words in each statement identified by them. This table of word-spans appears in Figure 13. Where word-spans are similar for two or more of the subjects (that is, where statement identification co-occurs), the co-occuring word-spans appear adjacent to one another. I have left gaps where they do not co-occur to make visual comparison easier.

As can be seen from Figure 13, the three subjects identified a different number of statements in the sample text, but not with great disparity. Subject One identified forty-two individual chunks of text as being statements; Subject Two identified forty-four and Subject Three identified forty-six. Although thirty-three statements were identified in common by all three subjects, thirty-nine were

FIGURE 13

CO-OCCURING WORD-SPANS FOR THREE SUBJECTS

SUBJECT NUMBER		
ONE	TWO	THREE
1-25	1-25	1-25
26-73	26-73	26-73
74-112	74-112	74-112
113-126	113-126	113-126
127-143	127-143	127-143
144-180	144-180	144-180
181-211	181-211	
		181-197
		198-211
212-236	212-236	212-236
	237-244	
	254-264	
237-264		
		237-264
265-300	265-300	264-300
301-310	301-310	301-310
311-348	311-348	311-348
349-362	349-362	
		349-353
		354-362
363-377	363-377	363-377
378-396	378-396	378-396
397-422	397-422	397-422
423-453	423-453	423-453
454-485	454-485	454-485
486-519	486-519	486-519
520-552		
	520-536	
		520-544
	537-552	
		545-552
553-575	553-575	553-575
576-627	576-627	576-627
628-677	628-677	628-677
678-699	678-699	678-699
700-705	700-705	
706-725	706-725	
		700-725
726-735	726-735	726-735
736-777	736-777	
		736-786
778-808	778-808	
		789-808
809-818	809-818	809-818
819-831	819-831	819-831
832-848	832-848	832-848
849-852	849-852	849-852
853-877	853-877	853-877
878-898	878-898	878-898
899-931	899-931	899-931
932-965	932-965	932-965
		966-971
966-983	966-983	
		972-978
		979-983
984-1013	984-1013	984-1013
1014-1032	1014-1032	1014-1032
1033-1066	1033-1066	1033-1066
1067-1079	1067-1079	1067-1079

identified by two or more subjects in common. The results also show that some 'chunks' of the text were seen as one statement, whilst the same 'chunks' were seen to be two statements by others. Graphs representing the distribution of statements identified by the three subjects in this experiment appear as Figures 14, 15 and 16.

Figure 17 shows a summary of those word-spans which have been similarly recognised classified by all three subjects. The word-span co-ordinates for all those chunks of text similarly identified by all subjects are given in the left-most column. Alongside each going from left-to-right across the table, are the code numbers as agreed by the subjects to these statements. The right-most two columns show the number of times each ⟨phase⟩ of the grammar has been used; for example, '3x1' means that three subjects used ⟨Phase One⟩ in their classification of the first statement in the table and '2x1A' means that two subjects used the code 1A which is ⟨Introduction⟩ in the grammar. In this sample, twelve out of thirty-three statements (36.36%) were assigned the same ⟨phase⟩ by all three subjects and twenty-eight (or 84.84%) were assigned the same ⟨phase⟩ by two or more subjects. The total given the same code-number by three subjects was three (9.09%) and by two or more candidates was fifteen (45.45%). As I have mentioned previously, the emphasis here is on the use of the three-phase structure, not the order of the grammar labels. Therefore these results suggest to me that the grammar can be used to classify statements in text and that the three-phase structure is an appropriate one for representing the concept of empirical argument formats in science text.

To attempt some statistical analysis of these results in order to demonstrate their significance, the following null hypothesis was set up and tested using a chi-square formula.

The null hypothesis is that each of the 3 subjects assigned each statement to one of the three phases from the grammar at random and independently. Therefore, the probability of a given statement being assigned to a given ⟨phase⟩ by a given subject was one third (1/3).

The probability thal all three subjects would agree on a given statement is equal to the probability that subjects 2 and 3 will choose the same ⟨phase⟩ as subject 1 - regardless of the statement chosen - is 1/3 x 1/3, which gives a probability of 1/9.

The probability that all three subjects would disagree on the same classification, equals the probability that (given the ⟨phase⟩ chosen by subject 1) subject 2 will choose one of the remaining 2 phases and subject 3 will choose the sole phase then remaining. So, probability $2/3 \times 1/3 = 2/9$. The probability that 2 out of 3 subjects will agree, is equal to what remains from the above. That is, 6/9 which equals a probability of 2/3.

So, from a total of 33 statements similarly identified by the 3 subjects the expected values for use in a chi-square test are:

1/9 x 33	$3\frac{2}{3}$
2/3 x 33	22
2/9 x 33	$7\frac{1}{3}$

Making a continuity correction to the above for the χ^2 test so,

$$\chi^2 = \sum_i \frac{(|O_i - E_i| - 1/2)^2}{E_i}$$

we get a value $\chi^2 = 18.5$, with 2 degrees of freedom which appears highly significant. However, the expected value for all phases the same is only $3\frac{2}{3}$, which is really too low to rely on χ^2. Amalgamation of two groups in whichever combination does not help to produce a more reliable result, because most of the difference between expected and observed values arises from transfers between the final two groups any way. If, however, the first and last groups were amalgamated thus,

	1st and Last	Middle
Expected	11	22
Observed	17	16

the continuity correction would be $\chi^2 = 4.9$, with one degree of freedom, which is just significant at the 5% level.

Therefore, although a larger sample may have produced better or more significant results, we can reject the null hypothesis here and show that the

observed values of 12, 16 and 5 statements which were similarly classified by the three subjects, can be shown as being significantly better than the expected values under the null hypothesis of 4, 22 and 7.

4.4.3 Conclusions from Part II of the Pilot Study

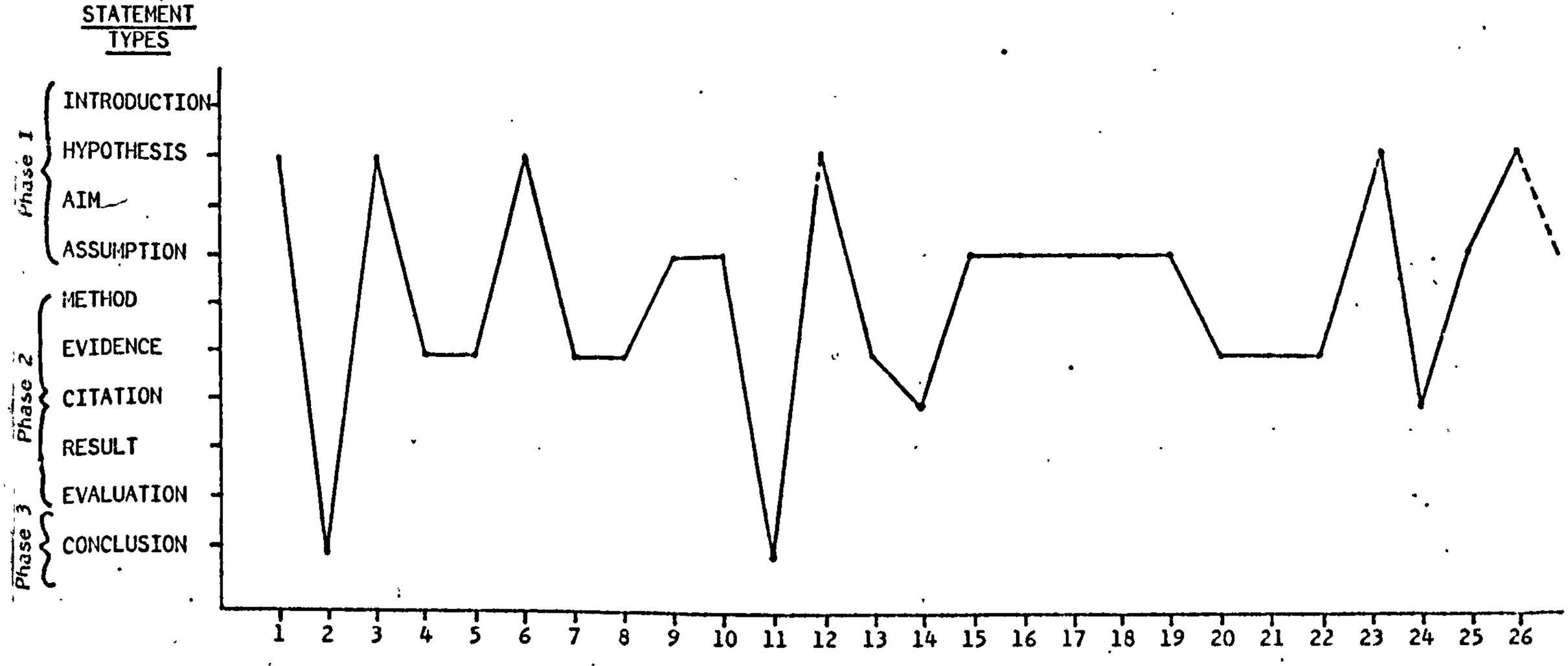
One further observation can be made from these results. It seems that a large number of chunks of text are recognised as being 'statements' by subjects. A closer examination of the structure and contents of those chunks which were similarly recognised by all three subjects could give some indications as to how we might develop algorithms for recognising statements automatically. Although this kind of analysis shall not take place here, I shall return to this topic when I discuss further work at the end of the thesis.

As shown above, the observed results of statement classification by subjects in this experiment were significantly better than those expected by the null hypothesis outlined. Such a results gives some weight to the contension that my grammar can be used as a tool for text analysis by humans and perhaps with appropriate rules, by computer. More particularly, the result indicates that the individual elements, or at least those used by all subjects for given statements, are appropriate semantic descriptors with which to classify given 'chunks' of text.

More data in the form of a large number of subjects analysing a greater number of texts would probably give a better basis for statistical analysis. The following experiments are an attempt to increase the number of subjects in two categories of experience with reading science text, in order to see how far the results from here can be extrapolated into a situation where more data will be available.

FIGURE 14

GRAPH OF STATEMENT-TYPE DISTRIBUTION FOR SUBJECT 1 IN PART II OF THE PILOT STUDY



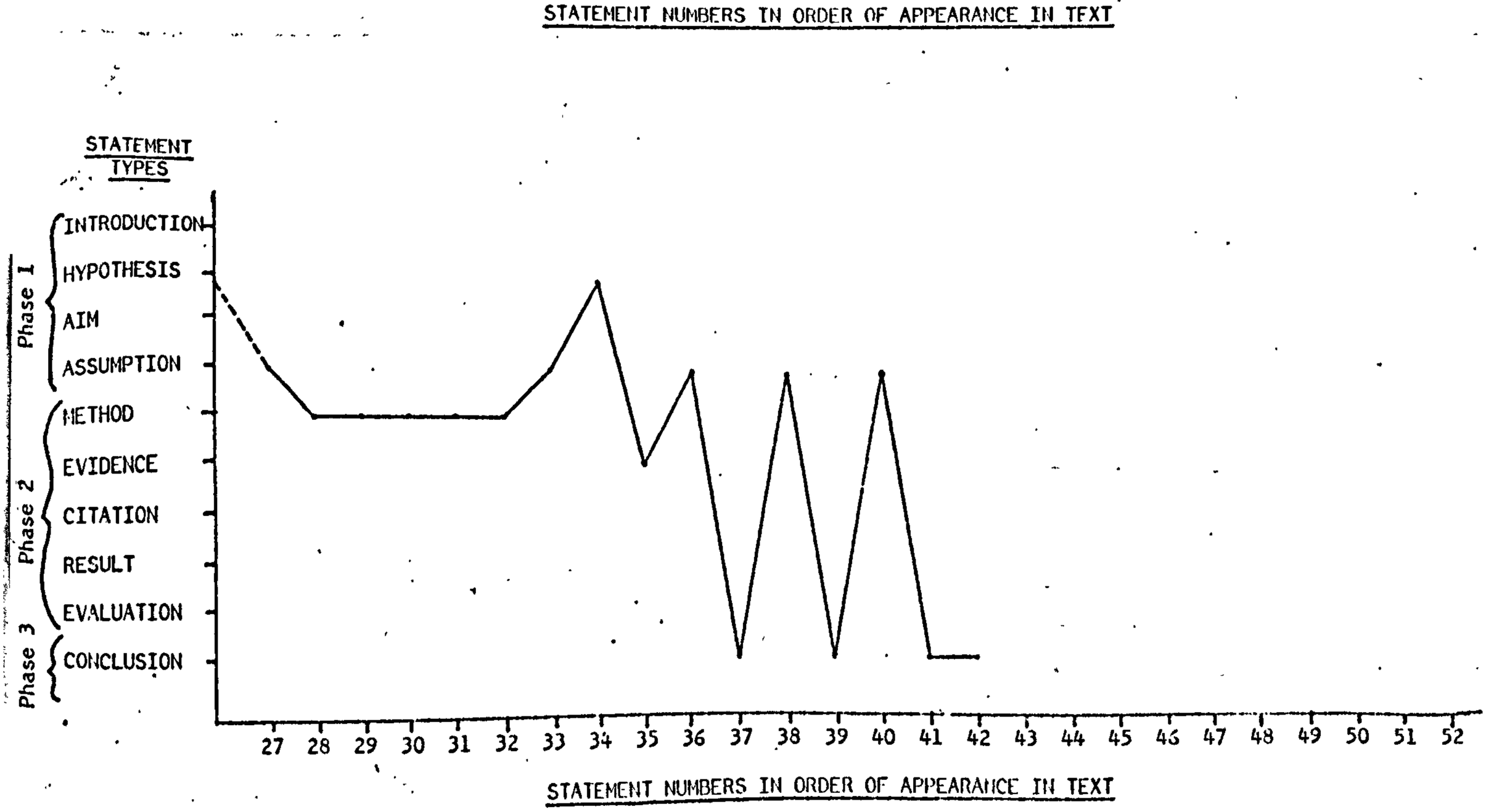


FIGURE 15

GRAPH OF STATEMENT-TYPE DISTRIBUTION FOR SUBJECT 2 IN PART II OF THE PILOT STUDY

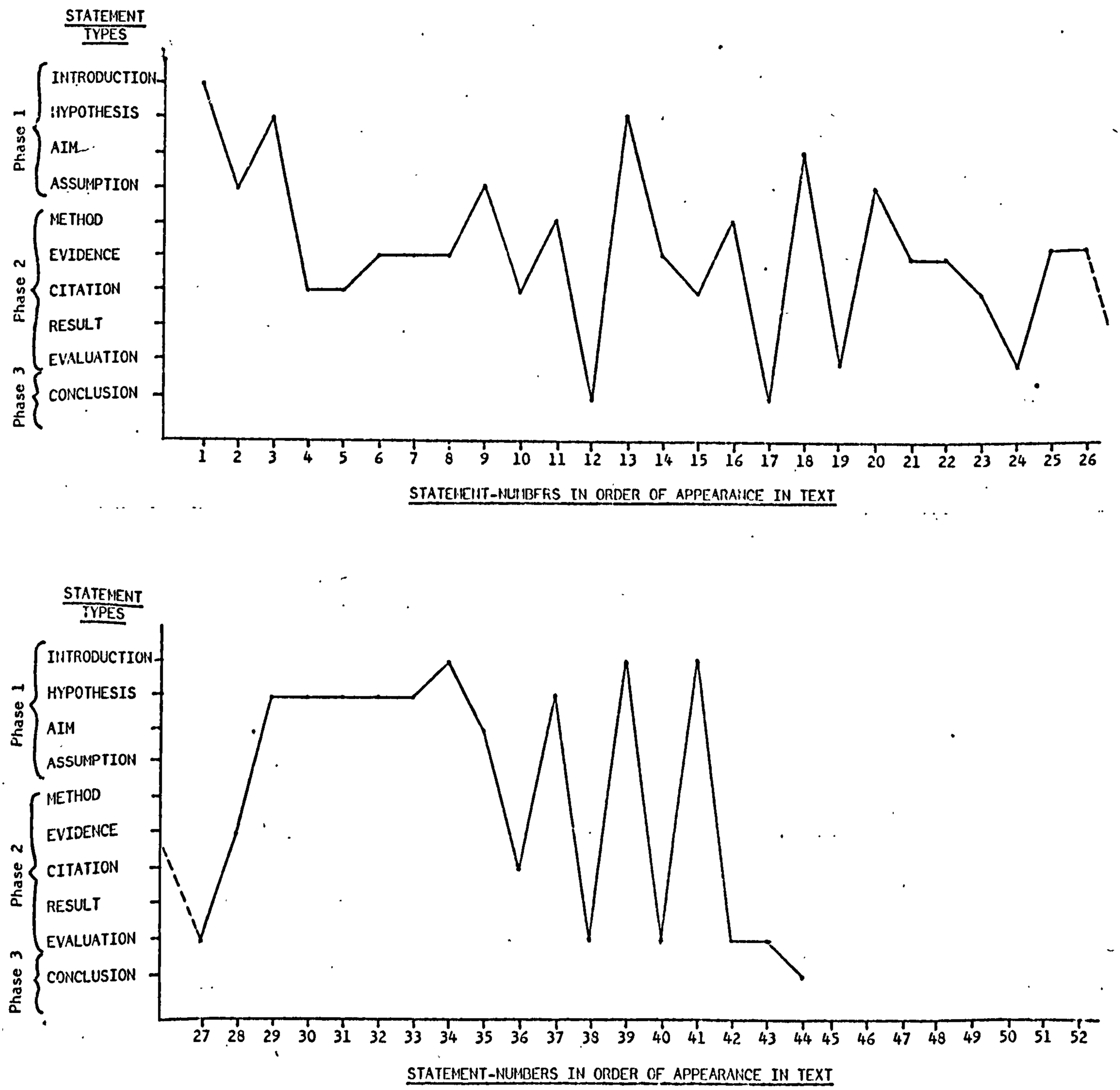
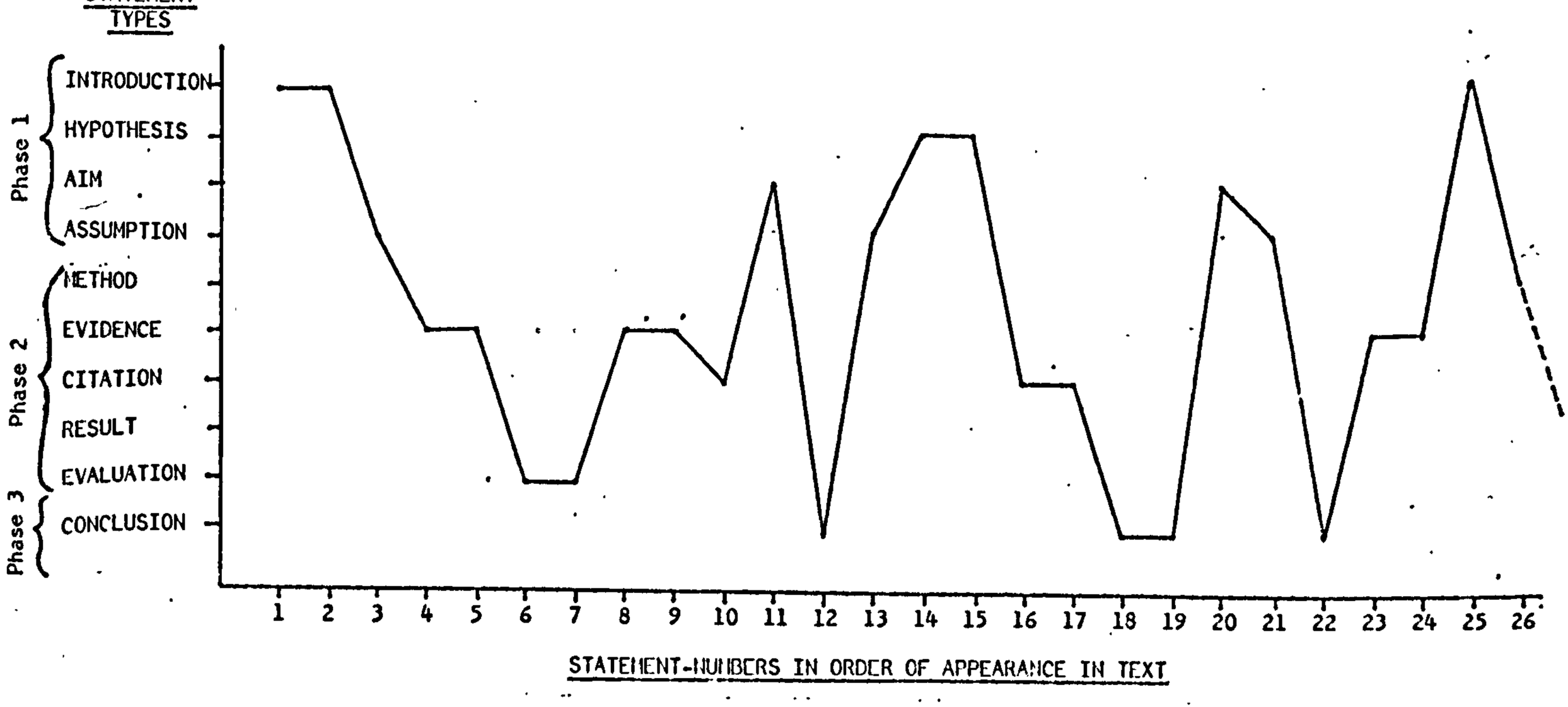


FIGURE 16

GRAPH OF STATEMENT-TYPE DISTRIBUTION FOR SUBJECT 3 IN PART II OF THE PILOT STUDY



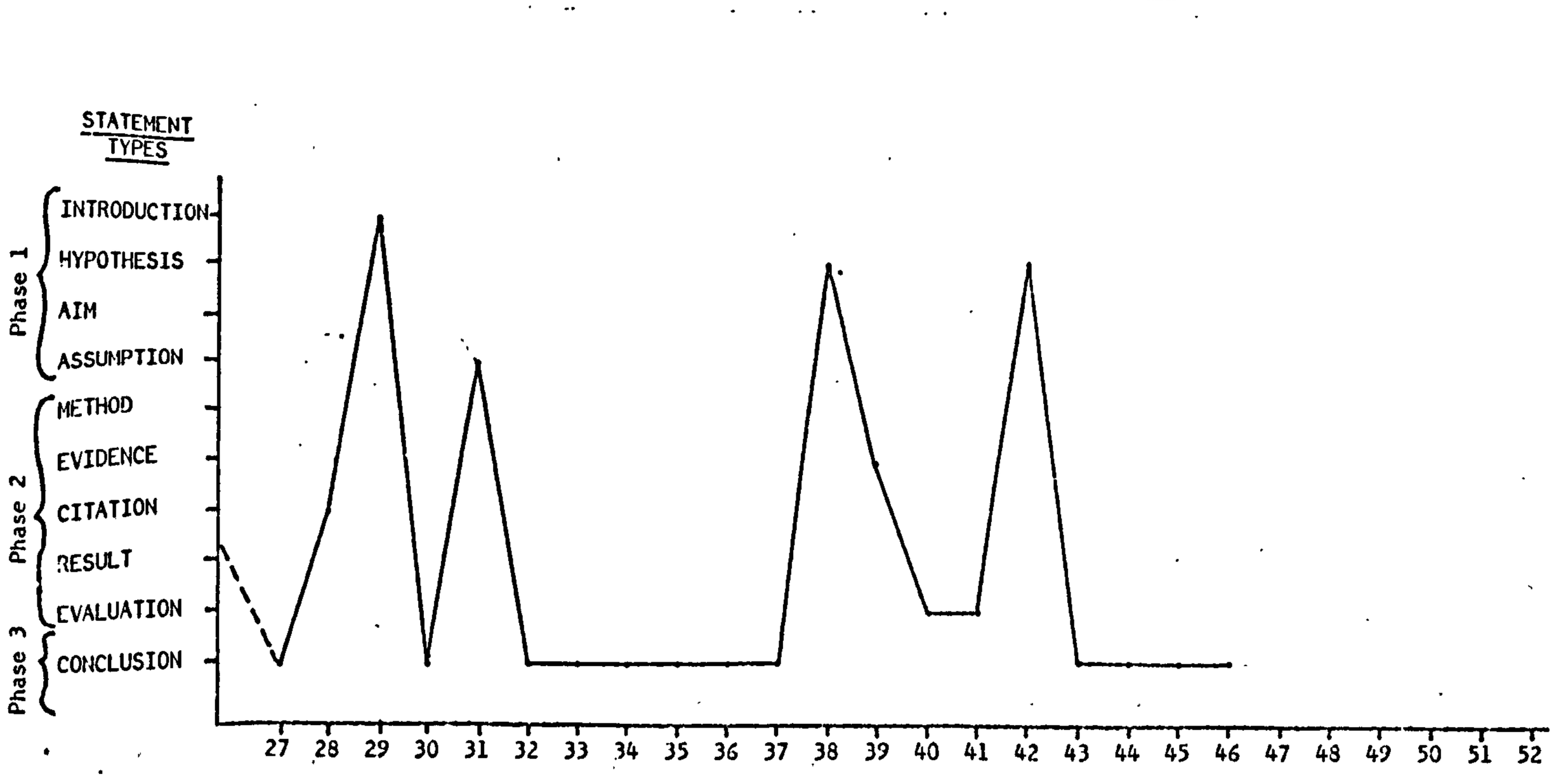


FIGURE 17

SUMMARY OF CO-OCCURRING STATEMENTS WITH CODES

WORD-SPAN CO-ORDINATES	SUBJECT-ONE CODES	SUBJECT TWO CODES	SUBJECT THREE CODES	CLASSIFICATION CORRELATIONS	
				Phase	Code
1-25	1B	1A	1A	3x1	2x1A
26-73	3A	1D	1A	2x1	0
74-112	1B	1B	1D	3x1	2x1B
113-126	2B	2C	2B	3x2	2x2B
127-143	2B	2C	2B	3x2	2x2B
144-180	1B	2D	2E	2x2	0
212-236	2B	2D	2B	3x2	2x2B
265-300	2D	2A	1C	2x2	0
301-310	3A	3A	3A	3x3	3x3A
311-348	1B	1B	1D	3x1	2x1B
363-377	2C	2C	2C	3x2	3x2C
378-396	1D	2A	3A	0	0
397-422	1D	1C	1C	3x1	3x1C
423-453	1D	2E	1D	2x1	2x1D
454-485	1D	1D	3A	2x1	2x1D
486-519	1D	2B	2B	2x2	2x2B
553-575	2B	2C	2A	3x2	0
576-627	2B	2E	3A	2x2	0
628-677	1B	2B	2C	2x2	0
678-699	2C	2B	1A	2x2	0
726-735	1B	2B	3A	0	0
809-818	2A	1B	3A	0	0
819-831	2A	1B	3A	0	0
832-848	2A	1B	3A	0	0
849-852	1D	1A	1B	3x1	0
853-877	1B	2C	2B	2x2	0
878-898	2B	2C	2E	3x2	0
899-931	1D	2A	2E	2x2	0
932-965	3A	2E	1B	0	0
984-1013	3A	2E	3A	2x3	2x3A
1014-1032	1D	1A	3A	2x1	0
1033-1066	3A	2E	3A	2x3	2x3A
1067-1079	3A	3A	3A	3x3	3x3A

Overall the results from this experiment seem to indicate that an analysis of a sample text by a large group of individuals using my grammar would be worthwhile. Part I of this Pilot Study showed that statement-types in particular sections of text rarely follow the semantic inference of the section heading. This experiment has shown how three subject specialists were able to classify statements using my grammar with a reasonable amount of commonality in choice of grammar elements. Moreover, they seemed to recognise a large proportion of statements in the text as being similar chunks of it. Each of the three participants in this experiment produced a 'non-linear' representation of the authors' argument. How much these trends are reflected in the results of two large groups of participants using another sample text can be seen in the following two experiments.

4.5 STATEMENT CLASSIFICATION BY SCIENTISTS

This experiment involved twenty-one subjects, all who were graduates in pure or applied science. The text used for the experiment was essentially non-technical but one which referred to an empirical investigation in Astronomy. A copy of the text, (1126 words long), can be seen in Appendix D. This was the kind of text I wanted to use for both this and the following experiment where 'non-scientists' would be reading the same text. I did not want either group to be overcome by technical description or symbolic representations such as complex mathematical proofs. I also chose this text for its absence of section headings so as not to prejudice the classification of individual statements. A set of instructions given to participants in this and the following experiment, together with the data recording sheets, can be seen after the sample text referred to above. In the instructions to participants can be seen directions for what to do if they find individual statements to be ambiguous or unclassifiable. This situation did not arise in the Pilot Study. For convenient reference, I have repeated below the coding structure for elements in the grammar and added the symbols for non-classifiable and ambiguous statements:

CODE	GRAMMAR ELEMENT
1A	Introduction
1B	Aim
1C	Hypothesis
1D	Assumption
1E	Observation
2A	Method
2B	Data (Evidence)
2C	Citation
2D	Result
2E	Evaluation
3A	Conclusion
?	Non-classifiable
*	Ambiguous

The experiment ran for approximately one hour. A few finished before that time and a few afterwards; most seemed to complete the task within the one hour period.

4.5.1 Analysing the results

The method used for analysing the results from this experiment was the same as that used for Part II of the Pilot Study. Each word in the sample text was given a number from 1 to 1126. Subjects were asked to place a sequential number in front of each statement they identified. This having been done I was able to transcribe word-spans for each subject onto a summary sheet. The word-spans on this summary sheet were then re-ordered to show the distribution of co-occurring 'chunks' of text which were identified by the 21 subjects. This summary can be seen as Figure 18. For instance, 20 out of 21 subjects identified the 'chunk' of text consisting of words 1-29 as being the first statement. Subject number 9 thought the first statement ranged from words 1-95. Subject number 3, 7, 12, 13 and 19 did not finish the task of numbering statements up to the last word.

FIGURE 18

CO-OCCURRING WORD-SPANS FOR 21 SCIENTISTS

SUBJECT NUMBER																				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29		1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29
								1-95												
		30-46								30-46										
			30-51				30-51				30-51	30-51					30-51	30-51	30-51	
30-95	30-95			30-95	30-95	30-95			30-95				30-95	30-95	30-95	30-95				30-95
								96-153												
		47-51								47-51										
			52-95				52-95					52-95					52-95			
		52-79									52-79							52-79	52-79	
96-117	96-117		96-117	96-117	96-117	96-117	96-117		96-117				96-117	96-117	96-117	96-117				96-117
								154-198												
										52-69										
										70-95										
		80-95									80-95							80-95	80-95	
								199-219												
		96-111								96-111	96-111	96-111					96-111	96-111	96-111	
118-153	118-153		118-153	118-153	118-153	118-153	118-153		118-153	118-153				118-153	118-153	118-153		118-153	118-153	118-153
		118-134									118-134	118-134	118-134				118-134			
154-198	154-198		154-198		154-198	154-198	154-198		154-198			154-198	154-198	154-198	154-198	154-198	154-198		154-198	154-198
				154-219						154-219										
220-236	220-236	220-236	220-236		220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236	220-236
		135-153									135-153	135-153	135-153				135-153			
		112-117								112-117	112-117	112-117					112-117	112-117	112-117	
199-219	119-219	119-219	119-219		119-219	119-219	119-219		119-219		119-219	119-219	119-219	119-219	119-219	119-219	119-219	119-219	119-219	119-219
				220-296																
	237-296					237-296		237-296					237-296							
297-313	297-313	297-313	297-313	297-313	297-313	297-313	297-313	297-313	297-313	297-313		297-313	297-313	297-313	297-313	297-313	297-313	297-313	297-313	
237-258			237-258		237-258		237-258		237-258	237-258	237-258			237-258	237-258	237-258	237-258			237-258
				314-417																
314-324	314-324	314-324	314-324		314-324	314-324	314-324	314-324	314-324		314-324	314-324			314-324	314-324		314-324	314-324	
		154-171									154-171							154-171		
259-296		259-296	259-296		259-296		259-296		259-296	259-296	259-296	259-296		259-296	259-296	259-296	259-296	259-296	259-296	259-296
418-460	418-460	418-460	418-460	418-460	418-460	418-460	418-460	418-460	418-460		418-460	418-460		418-460	418-460	418-460	418-460		418-460	418-460
	325-356	325-365	325-365		325-365	325-365	325-365	325-365	325-365		325-365	325-365			325-365	325-365			325-365	325-365
		172-198									172-198							172-198		
				416-502		416-502														

FIGURE 1.8 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	366-417		366-417		366-417	366-417	366-417	366-417	366-417					366-417	366-417	366-417				366-417
																				297-324
				508-549																
		237-250										237-250						237-250	237-250	
													314-365	314-365						
325-357																		325-357		
				550-602																
		251-258										251-258						251-258	251-258	
													366-502							
358-365																	358-365	358-365		
				603-637							603-637									
484-502		484-502	484-502		484-502		484-502	484-502	484-502		484-502	464-502		484-502	484-502	484-502				
503-514	503-514	503-514	503-514			502-514	503-514		503-514		503-514	503-514	503-514	503-514				503-514	503-514	
366-404																			366-404	
	461-488																			
638-666	638-666	638-666		638-666	638-666	638-666	638-666	638-666	638-666	638-666				638-666	638-666	638-666	638-666			638-666
					503-523			503-523							503-523	503-523	503-523			
										314-335										
													515-549						515-549	
																	314-357			
																				461-523
461-483		461-483	461-483		461-483		461-483	461-483	461-483		461-483	461-483		461-483	461-483	461-483				
405-417																				
	489-502																			
667-683	667-683	667-683	667-683	667-683	667-683	667-683	667-683	667-683	667-683		667-683	667-683		667-683	667-683	667-683	667-683		667-683	667-683
	524-549				524-549	524-549	524-549	524-549	524-549					524-549	524-549	524-549	524-549	524-549		524-549
										336-344										
											297-306									
													550-637							550-637
	684-760			684-760	684-760	684-760														
515-523	515-523	515-523	515-523			515-523	515-523		515-523	515-523	515-523	515-523		515-523				515-523		
	550-574							550-574						550-574	550-574	550-574				
										345-365										
											307-313									
													638-729							
																	365-378			
	761-729			761-729	761-729	761-729														
								575-637												
										366-409		366-409								
	730-770			730-770									730-770				730-770			
																	379-417			

FIGURE 18 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
		550-561				550-561				550-561										
		410-417								410-417		410-417								
771-792	771-792		771-792		771-792	771-792	771-792	771-792	771-792		771-792	771-792	771-792	771-792	771-792	771-792	771-792		771-792	
																			405-417	
				771-817																
550-574-	550-574-		550-574-		550-574-		550-574-		550-574-		550-574-	550-574-					550-574-	550-574-	550-574-	
						562-574				562-574										
									793-817				793 817							
																	461-502			
																		366-424		
684-710							684-710	684-710	684-710		684-710	684-710		684-710	684-710	684-710	684-710		684-710	684-710
	575-602	575-602	575-602		575-602	575-602	575-602		575-602		575-602	575-602		575-602	575-602	575-602			575-602	
		366-378																		
524-535-		524-535	524-535							524-535										
				818-899																
										461-514										
													818-925							
																		425-460		
																		461-468	461-468	
								711-770												711-770
603-637	603-637	603-637	603-637		603-637	603-637	603-637		603-637			603-637		603-637	603-637	603-637				•
		379-404																		
536-549		536-549	536-549							536-549										
				890-925																
													926-987							
																			469-483	
																				771-828
		405-409																		
926-941	926-941			926-941	926-941		926-941	926-941	926-941			926-941		926-941	926-941	926-941	926-941		926-941	
													988-1034							
																		469-502		
																			484-490	
																				829-925
				942-987				942-987												
793-806			793-806			793-806		793-806		793-806	793-806	793-806		793-806	793-806	793-806			793-806	
	1035-1082												1035-1082							1035-1082
																	575-637			•
																			491 502	
																				926-1003
575-597										575-597								575-597		
968-1003			988-1003	988-1003	988-1003		988-1003	988-1003	988-1003			988-1003		988-1003	988-1003	988-1003	988-1003		988-1003	

FIGURE 18 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
807-817						807-817		807-817		807-817	807-817	807-817		807-817	807-817	807-817			807-817	
											524-549	524-549								
1083-1126				1083-1126	1083-1126		1083-1126	1083-1126	1083-1126				1083-1126		1083-1126		1083-1126			1083-1126
1004-1023				1004-1023	1004-1023		1004-1023	1004-1023	1004-1023					1004-1023	1004-1023	1004-1023	1004-1023		1004-1023	1004-1023
598-602																				
			638-642									638-642							638-642	
				1004-1034																
								818-844			818-844									
711-729		711-729	711-729				711-729		711-729	711-729	711-729	711-729		711-729	711-729	711-729			711-729	
1024-1034	1024-1034		1024-1034		1024-1034		1024-1034	1024-1034	1024-1034	1024-1034				1024-1034	1024-1034	1024-1034			1024-1034	1024-1034
			643-647									643-647							643-647	
				1035-1068																
730-755			730-755		730-755	730-755	730-755		730-755	730-755		730-755		730-755		730-755			730-755	
845-862					845-862	845-862	845-862	845-862	845-862					845-862	845-862	845-862			845-862	
															730-770					
			648-654									648-654							648-654	
				1069-1082				1069-1082									1069-1082			
756-770			756-770		756-770	756-770	756-770		756-770	756-770		756-770		756-770		756-770			756-770	
863-889					863-889			863-889						863-889	863-889	863-889				
										598-637									598-637	
	793-844																			
			655-660									655-660							655-660	
890-907	890-907				890-907	890-907	890-907	890-907	890-907	890-907				890-907	890-907	890-907			890-907	
	845-859																			
			661-666									661-666							661-666	
					793-817		793-817													
908-925	908-925		908-925		908-925	908-925	908-925	908-925		908-925		908-925		908-925	908-925	908-925			908-925	
										667-710										
818-828					818-828	818-828	818-826		818-828			818-828		818-828	818-828	818-828			818-828	
																	793-925			
		684-694	684-694																	
829-844					829-844	829-844	829-844		829-844			829-844		829-844	829-844	829-844			829-844	
		562-574																		
		695-710	695-710																	
	942-1004																			
			863-870			863-870	863-870					863-870							863-870	
									863-875	863-875	863-875									
											730-792									
978-987			978-987		978-987		978-987		978-987			978-987		978-987	978-987	978-987	978-987		978-987	
	1005-1023																			
																	1023-1034			

FIGURE 18 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
						871-889	871-859												871-889	
1035-1068					1035-1068		1035-1068	1035-1068	1035-1068					1035-1068	1035-1068	1035-1068	1035-1068		1035-1068	
									876-889	876-889										
									908-918											
										818-835										
	1083-1101		1083-1101							1083-1101				1083-1101		1083-1101			1083-1101	
			836-844							836-844										
			845-850							845-850	845-850	845-850								
942-977			942-977		942-977		942-977		942-977	942-977				942-977	942-977	942-977	942-977		942-977	
	1102-1126		1102-1126											1102-1126						
			807-828																	
									919-925											
			851-862							851-862	851-862	851-862								
			829-835																	
		730-741																		
		742-755																		
1069-1082			1069-1082		1069-1082					1069-1082					1069-1082				1069-1082	
			671-875									871-875								
			926-932							926-932										
							1069-1082		1069-1082					1069-1082		1069-1082				
			876-889									876-889								
			933-941							933-941										
			890-907									890-907								
										978-1003										
										1113-1126						1113-1126			1113-1126	
										1004-1012		1004-1012								
										1013-1023										
			1035-1053							1035-1053										
			1054-1068							1054-1068										
										1102-1112						1102-1112			1102-1112	
												942-958								
												959-977								
												1013-1034								

I now decided to rank all those results which gave word-spans with occurrences of greater than nine subjects. There is no statistical significance associated with this figure. Ten and above occurrences from a maximum of sixteen seems intuitively a reasonable level for comparison with the results from the next experiment with non-scientists. As can be seen in Figure 19, two word-spans were similarly identified by fifteen subjects; two word-spans by fourteen subjects; six word-spans by eleven subjects; and eight word-spans by ten subjects.

Following the ranking of word-spans in Figure 19, there is a table of codes assigned by each subject to each word-span. This is Figure 20. In the right-most column I have given summations for the most often occurring ⟨phase⟩ and ⟨code⟩ for each word-span. This table will be compared with the results from the next experiment after they have been presented.

4.5.2 Conclusions from this experiment

There are two major reasons why the results from this experiment cannot be compared with those from the Pilot Study. First the two texts used are different; second, the numbers of subjects used in the second experiment far exceeds those in the Pilot Study, thereby significantly altering the random probability of subjects choosing one or another grammar elements from the given set. Overall performance can be considered however, without attempting to apply any statistical measures to the results. For instance, even though the number of subjects was four times greater than for the Pilot Study, not one of the sixteen individuals identified a single word-span which concurred with that of any one of the other fifteen subjects. In fact, there were only two word-spans which were similarly identified by fifteen of the subjects.

This compares with twelve word-spans similarly identified by all three subjects in the Pilot Study and twenty-eight word-spans by two of the subjects. The texts had similar length; 1079 words in the Pilot Study and 1126 in this experiment. I would have assumed that this result would be reversed to some extent. That is, the larger the number of subjects, the more agreement there would be of classification codes for individual statements. As I have stated though, we are not comparing experiments with the same text here, so even that intuitive assumption cannot be tested. A more meaningful comparison can be made between the results of the next experiment with those from here.

Some word-spans were identified by a high proportion of the subjects as being classifiable using either ⟨phase⟩ or a descriptor. In some cases there is more agreement about which ⟨phase⟩ should be assigned, rather than which actual semantic label. In Figure 20 for example, word-span 1-29 was assigned the same ⟨phase⟩ by fifteen subjects and nine of those chose ⟨observation⟩as the precise semantic descriptor. Nine out of sixteen overall, or nine out of the fifteen who classified the same word-span is statistically significant in the crude sense that the result is greater than 50% for statement classification by subjects. A review of Figure 20 however, shows that whilst the correlations for statement classification for ⟨phase⟩ are high, they are only significant for initial statement recognition in the range from ten to fifteen occurrences. We must conclude from that then, that although subjects could use my grammar to classify statements in this text, there was not enough agreement on the initial statement identification to provide highly significant results. The results do show that the author has produced a text which although it contains 'semantically loaded' statements which from the grammar definition indicate method,

FIGURE 19

WORD-SPAN RANKS FOR ALL CO-OCCURRENCES GREATER THAN NINE FOR SIXTEEN SCIENTISTS

SPAN	OCCURRENCE	SUBJECT-NUMBER
1-29	15	1, 2, 4, 5, 6, 8, 10, 11, 14, 15, 16, 17, 18, 20, 21
220-236		1, 2, 4, 6, 8, 9, 10, 11, 14, 15, 16, 17, 18, 20, 21
418-460	14	1, 2, 4, 5, 6, 8, 9, 10, 15, 16, 17, 18, 20, 21
667-683		1, 2, 4, 5, 6, 8, 9, 10, 15, 16, 17, 18, 20, 21
118-153	13	1, 2, 4, 5, 6, 8, 10, 11, 15, 16, 17, 20, 21
199-219		1, 2, 4, 6, 8, 10, 14, 15, 16, 17, 18, 20, 21
154-198		1, 2, 4, 6, 8, 10, 14, 15, 16, 17, 18, 20, 21
638-666		1, 2, 5, 6, 8, 9, 10, 11, 15, 16, 17, 18, 21
771-792		1, 2, 4, 6, 8, 9, 10, 14, 15, 16, 17, 18, 20
1024-1034		1, 2, 4, 6, 8, 9, 10, 11, 15, 16, 17, 20, 21
96-117	12	1, 2, 4, 5, 6, 8, 10, 14, 15, 16, 17, 21
259-296		1, 4, 6, 8, 10, 11, 15, 16, 17, 18, 20, 21
926-941		1, 2, 5, 6, 8, 9, 10, 15, 16, 17, 18, 20
988-1003		1, 4, 5, 6, 8, 9, 10, 15, 16, 17, 18, 20
1004-1023		1, 5, 6, 8, 9, 10, 15, 16, 17, 18, 20, 21
237-258	11	1, 4, 6, 8, 10, 11, 15, 16, 17, 18, 21
890-907		1, 2, 6, 8, 9, 10, 11, 15, 16, 17, 20
908-925		1, 2, 4, 6, 8, 9, 11, 15, 16, 17, 20
942-977		1, 4, 6, 8, 10, 11, 15, 16, 17, 18, 20
30-95	10	1, 2, 5, 6, 10, 14, 15, 16, 17, 21
314-324		1, 2, 4, 6, 8, 9, 10, 16, 17, 20
325-365		2, 4, 6, 8, 9, 10, 16, 17, 20, 21
366-417		2, 4, 6, 8, 9, 10, 15, 16, 17, 21
524-549		2, 6, 8, 9, 10, 15, 16, 17, 18, 21
1083-1126		1, 5, 6, 8, 9, 10, 14, 16, 18, 21
978-987		1, 4, 6, 8, 10, 15, 16, 17, 18, 20
1035-1068		1, 6, 8, 9, 10, 15, 16, 17, 18, 20

FIGURE 20

COMPARING CODES FOR WORD-SPANS FROM SCIENTISTS

For Fifteen Occurrences

Word-span	SUBJECT-NUMBER 1	2	4	5	6	8	10	11	14	15	16	17	18	20	21	CORRELATIONS Phase	Code
1-29	1E	1E	1E	1E	1A	1E	1A	1E	1E	1A	1E	1E	1A	1A	1A	15x1	9x1E
Word-span	1	2	4	6	8	9	10	11	14	15	16	17	18	20	21		
220-236	2A	1B	2A	2A	2B	2A	2A	1C	2B	2B	2A	2A	1E	2B	2A	12x2	8x2A

For Fourteen Occurrences

Word-span	1	2	4	5	6	8	9	10	15	16	17	18	20	21			
418-460	2D	2A	2A	2A	2A	2A	1D	2A	2A	2A	2A	2A	2D	2A		13x2	11x2A
Word-span	1	2	4	5	6	8	9	10	15	16	17	18	20	21			
667-683	2E	2E	1E	2D	2E	3A	1C	2E	2E	3A	2E	2E	2E	3A		9x2	8x2E

For Thirteen Occurrences

Word-span	1	2	4	5	6	8	10	11	15	16	17	20	21				
118-153	1A	1A	1B	1A	1A	2B	1A	2B	1D	1E	1C	1B	1A			11x1	6x1A
Word-span	1	2	4	6	8	10	14	15	16	17	18	20	21				
199-219	2A	3A	3A	1C	3A	1C	1C	1C	1C	3A	1C	1C	3A			7x1	7x1C
Word-span	1	2	4	6	8	10	14	15	16	17	18	20	21				
154-198	1E	1D	1C	1C	2B	1E	1D	1D	1E	1D	1D	1D	1D			12x1	7x1D
Word-span	1	2	5	6	8	9	10	11	15	16	17	18	21				
638-666	2E	2D	2A	2D	2D	1C	2D	3A	2D	2D	2D	2D	2D			11x2	9x2D
Word-span	1	2	4	6	8	9	10	14	15	16	17	18	20				
771-792	1E	1C	1E	2E	2C	2E	2E	1E	2E	1E	1D	2E	2D			7x2	5x2E
Word-span	1	2	4	6	8	9	10	11	15	16	17	20	21				
1024-1034	3A	3A	2D	2E	1E	3A	2A	1C	2E	2A	2E	2D	3A			6x2	4x3A

For Twelve Occurrences

Word-span	1	2	4	5	6	8	10	14	15	16	17	21					
96-117	1C	3A	1E	1E	1A	1E	1E	1E	1E	1E	1C	1E				11x1	8x1E
Word-span	1	4	6	8	10	11	15	16	17	18	20	21					
259-296	2A	2C	2A	2A	2B	2A	2B	2B	2A	1E	2D	2A				11x2	6x2A
Word-span	1	2	5	6	8	9	10	15	16	17	18	20					
926-941	2A	1B	2E	2E	1A	3A	2D	2E	2A	1E	1E	2E				7x2	4x2E
Word-span	1	4	5	6	8	9	10	15	16	17	18	20					
988-1003	3A	1E	1E	2E	1E	3A	2B	2E	1E	1E	3A	1E				6x1	6x1E
Word-span	1	4	6	8	9	10	15	16	17	18	20	21					
1004-1023	3A	2A	2E	2A	2A	1C	2E	1D	2A	2A	2A	2A				9x2	7x2A

FIGURE 20 CONTINUED

For Eleven Occurrences

Word-span	1	4	6	8	10	11	15	16	17	18	21						
237-258	2B	2A	2A	2A	2B	2A	2B	2B	2A	1D	2A					10x2	6x2A
Word-span	1	2	6	8	9	10	11	15	16	17	20						
890-907	2A	1D	2E	2C	2A	2B	1E	2E	1E	1D	2D					7x2	2x2A
Word-span	1	2	4	6	8	9	11	15	16	17	20						
908-925	2A	3A	3A	2E	2D	2E	1E	2E	1D	2A	2E					7x2	4x2E
Word-span	1	4	6	8	10	11	15	16	17	18	20						
942-977	2B	1C	2E	1E	2A	1E	2E	1E	2B	2B	3A					6x2	3x2B

For Ten Occurrences

Word-span	1	2	5	6	10	14	15	16	17	21							
30-95	1E	1E	1E	1A	1A	1E	1E	1D	1E	1A						10x1	6x1E
Word-span	1	2	4	6	8	9	10	16	17	20							
314-324	1C	1E	1E	2B	3A	1E	2A	1E	1E	1D						7x1	5x1E
Word-span	2	4	6	8	9	10	16	17	20	21							
325-365	1D	1D	1D	1D	2A	1D	1D	1D	2A	1D						8x1	8x1D
Word-span	2	4	6	8	9	10	15	16	17	21							
366-417	2A	1E	1D	1E	2D	2E	1D	1D	2D	2A						5x17 5x25	3x1D
Word-span	2	6	8	9	10	15	16	17	18	21							
524-549	1E	1C	2E	2D	2B	2B	1E	1C	2B	1C						5x17 5x25	3x2B 3x1C
Word-span	1	5	6	8	9	10	14	16	18	21							
1083-1126	3A	3A	2E	3A	3A	3A	3A	3A	3A	3A						9x3	9x3A
Word-span	1	4	6	8	10	15	16	17	18	20							
978-987	1E	1D	2E	1E	1D	2E	2E	1D	3A	2E						5x1	4x2E
Word-span	1	6	8	9	10	15	16	17	18	20							
1035-1068	3A	2E	2E	3A	2D	3A	2E	3A	3A	3A						6x3	6x3A

results and so on, essentially has a non-linear argument, in terms of the 'conventional format'.

4.6 STATEMENT CLASSIFICATION OF NON-SCIENTISTS

The subjects for this experiment were all experienced indexers and abstractors for scientific and technical literature. They were given the same experimental conditions as the scientists in the previous group and asked to perform the same task using the same text. The results from this experiment are presented in the same manner as those in the previous experiment and a direct comparison can be made between one table and another in the two sections of results.

4.6.1 Analysing the results

Once again, the results from this experiment have been presented in tabular form showing word-spans for each identified 'chunk' of text by individual subjects. The tables which follow (Figures 21 and 22) are self-explanatory and should be read in a similar way to the tables in the previous section. In this experiment only nine of the twenty-one subjects completed the task compared with fifteen in the previous experiment. The ranked word-spans for these nine subjects have been selected from those which have greater than five occurrences. This is compared with a ranked sequence of word-spans for fifteen subjects which had greater than nine occurrences in the previous experiment.

4.6.2 Conclusions from this experiment

There were no word-spans which were similarly identified by all twenty-one subjects, but of the nine subjects who did complete the task, the largest number of co-occurrences were eight - this result for only two word-spans. Four word-spans were similarly identified by seven subjects and eight by the remaining six subjects. As can be seen in Figure 23, the instances when ⟨phases⟩ are similarly used

FIGURE 21

CO-OCCURRING WORD-SPANS FOR 21 NON-SCIENTISTS

SUBJECT NUMBER																				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1-29	1-29	1-29		1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29	1-29		1-29	1-29		1-29	1-29	1-29	1-29
			1-153										1-153			1-153				
30-95	30-95	30-95		30-95	30-95	30-95	30-95	30-95	30-95	30-95	30-95			30-95	30-95			30-95	30-95	30-95
																	30-51			
												30-46								
154-198	154-198	154-198		154-198				154-198	154-198	154-198	154-198		154-198	154-198	154-198	154-198	154-198	154-198		154-198
			154-219		154-219	154-219														
96-117	96-117	96-117		96-117	96-117			96-117	96-117	96-117	96-117	96-117		96-117	96-117			96-117	96-117	96-117
												52-95					52-95			
199-219	199-219	199-219			199-219			199-219	199-219	199-219	199-219	199-219	199-219	199-219	199-219	199-219	199-219	199-219	199-219	199-219
												47-51								
							96-111										96-111			
						96-134														
220-313	220-313		220-313							220-313			220-313			220-313				
118-153	118-153	118-153		118-153	118-153	118-153		118-153		118-153				118-153	118-153			118-153		118-153
			314-502										314-502			314-502				
							112-117												112-117	
							118-134		118-134		118-134	118-134							118-134	
			503-549							503-549			503-549							
									135-153		135-153	135-153					135-153		135-153	
																	112-134			
550-637			550-637		550-637	550-637				550-637			550-637			550-637				
					220-296				220-296											
		220-236		220-236		220-236	220-236	220-236			220-236	220-236		220-236	220-236			220-236	220-236	220-236
							135-140													
		503-514	503-514				503-514	503-514			503-514	503-514			503-514	503-514	503-514	503-514	503-514	503-514
				297-313	297-313						297-313	297-313						297-313	297-313	297-313
						237-296						237-296								
							141-153													
		515-523		515-523			515-523	515-523			515-523	515-523			515-523	515-523	515-523	515-523	515-523	515-523
							154-171												154-171	
314-365	314-365				314-365					314 365										
		237-258		237-258				237-258			237-258			237-258	237-258			237-258	237-258	237-258
			667-729										667-729			667-729				
		297-313				297-313	297-313	297-313	297-313						297-313		297-313			
												154-159								
524-549		524-549		524-549	524-549	524-549	524-549	524-549	524-549		524-549	524-549			524-549	524-549	524-549		524-549	524-549

FIGURE 21 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
												172-198							172-198	
366-417							366-417				366-417	366-417			366-417		366-417	366-417		366-417
		259-296		259-296				259-296			259-296			259-296	259-296		259-296	259-296	259-296	259-296
			730-770							730-770			730-770			730- 770				
		366-417			366-417			366-417	366-417	366-417									366-417	
		314-324		314-324		314-324		314-324	314-324		314-324				314-324		314-324	314-324	314-324	314-324
							172-219													
												160-171								
																	220-258			
418-460		418-460		418-460	418-460		418-460	418-460	418-460	418-460	418-460	418-460			418-460		418-460	418-460	418-460	418-460
			771-925																	
		325-365				325-365	325-365	325-365	325-365		325-365				325-365		325-365	325-365	325-365	325-365
461-483		461-483			461-483			461-483			461-483	461-483			461-483			461-483		461-483
			926-1034							926-1034									926-1034	
						366-404														
							237-250													
						461-502	461-502		461-502	461-502							461-502		461-502	
					771-817								771-817							
484-502		484-502		484-502	484-502			484-502			484-502	484-502			484-502			484-502		484-502
			1035-1082															1035-1082	1035-1082	
				325-357																
						405-460														
							251-296													
													818-925							
503-523					503-523	503-523				503-523										
			1083-1126															1083-1126	1083-1126	
				358-378																
													926-977							
				379-417																
							314-324													
638-666			638-666	638-666	638-666				638-666	638-666		638-666	638-666		638-666	638-666	638-666	638-666		638-666
													978-1034							
										667-710										
												314-335								
													1035-1126		1035-1126					
				461-468																
				711-729			711-729	711-729	711-729	711-729	711-729	711-729			711-729		711-729	711-729	711-729	711-729
												336-365								
				469-483																
				667-683	667-683		667-683	667-683	667-683		667-683	667-683			667-683		667-683	667-683	667-683	667-683
				550-561					550-561			550-561								
					684-729															

FIGURE 21 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
				562-574					562-574			562-574								
										771-828										
				730-755	730-755		730-755	730-755				730-755			730-755				730-755	
				575-602			575-602	575-602	575-602			575-602			575-602		575-602	575-602	575-602	575-602
										829-862								829-862		
				756-770	756-770		756-770	756-770				756-770			756-770				756-770	
							550-574	550-574			550-574				550-574		550-574	550-574	550-574	550-574
				603-637				603-637	603-637		603-637								603-637	
										863-925								863-925		
																		536-549		
				818-828	818-828		818-828	818-828				818-828					818-828	818-828	818-828	
				1035-1068				1035-1068		1035-1068							1035-1068	1035-1068		
											575-588									
							603-637					603-637			603-637		603-637	603-637	603-637	603-637
							638-647	638-647			638-647								636-647	
									684-694											
										1069-1126										
											589-597									
							648-654	648-654			648-654								646-654	
									695-710											
											598-602									
							655-660	655-660			655-660								655-660	
				684-710			684-710	684-710			684-710	684-710			684-710		684-710	684-710	684-710	684-710
							661-666	661-666			661-666								661-666	
									730-741											
									742-755											
																	730-770	730-770		
				771-792			771-792	771-792				771-792					771-792	771-792	771-792	
															771-1034					
				793-817													793-817	793-817		
																	829-844			
							845-862					845-862					845-862		845-862	
							793-806	793-806				793-806							793-806	
							863-875	863-875									863-875		863-875	
																		926-987		
							876-889										876-889			
							807-817	807-817				807-817							807-817	
				829-907																
																	890-925			
																		988-1034		
				908-925			908-925	908-925											908-925	
								829-835												

FIGURE 21 CONTINUED

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
				926-941				926-941										926-941		
																		942-977		
				988-1023																
							836-844												836-844	
							890-907	890-907												
																	1004-1023			
				1024-1034				1024-1034									1024-1034			
				1069-1082													1069-1082			
								942-1023												
																			876 907	
				1083-1101				1083-1101									1083-1101			
				1102-1112				1102-1112									1102-1112			
				1113-1126				1113-1126												

FIGURE 22

WORD-SPAN RANKS FOR ALL OCCURRENCES GREATER THAN FIVE FOR NINE NON-SCIENTISTS

SPAN	OCCURRENCE	SUBJECT-NUMBER
199-219	8	5, 9, 11, 14, 16, 18, 19, 20
1-29	7	5, 9, 11, 16, 18, 19, 20
154-198		5, 9, 11, 14, 16, 18, 19
418-460		5, 9, 11, 16, 18, 19, 20
711-729		5, 9, 11, 16, 18, 19, 20
30-95	6	5, 9, 11, 16, 19, 20
515-523		5, 9, 16, 18, 19, 20
259-296		5, 9, 16, 18, 19, 20
314-324		5, 9, 16, 18, 19, 20
638-666		5, 11, 14, 16, 18, 19
667-683		5, 9, 16, 18, 19, 20
575-602		5, 9, 16, 18, 19, 20
684-710		5, 9, 16, 18, 19, 20

FIGURE 23

COMPARING CODES FOR WORD-SPANS FROM NON-SCIENTISTS

For Eight Occurrences

	SUBJECT-NUMBER								CORRELATIONS	
Word-span	5	9	11	14	16	18	19	20	Phase	Code
119-219	1C	2E	1D	3A	1D	1C	1E	2A	5x1	2x1E 2x1D

For Seven Occurrences

Word-span	5	9	11	16	18	19	20			
1-29	1E	1E	1A	1A	1E	1A	1A		7x1	4x1A
Word-span	5	9	11	14	16	18	19			
154-198	2A	1D	2B	1C	1C	1E	1E		5x1	2x1E 2x1C
Word-span	5	9	11	16	18	19	20			
418-460	2E	2D	2A	2A	2A	1D	2D		6x2	3x2A
Word-span	5	9	11	16	18	19	20			
711-729	1B	2E	2A	2D	2A	2D	2D		6x2	3x2D

For Six Occurrences

Word-span	5	9	11	16	19	20				
30-95	1E	1E	1E	1A	1E	1C			6x1	4x1E
Word-span	5	9	16	18	19	20				
515-523	1B	1C	2B	1A	2E	1E			4x1	0
Word-span	5	9	16	18	19	20				
259-296	2A	2D	2A	2A	2D	2B			6x2	3x2A
Word-span	5	9	16	18	19	20				
314-324	2E	1D	1D	2E	2A	2A			4x2	2x2A 2x2E
Word-span	5	11	14	16	18	19				
638-666	2C	2D	2D	2D	1E	2D			5x1	4x2D
Word-span	5	9	16	18	19	20				
667-683	2D	3A	2E	1E	2D	1E			3x2	2x2D 2x1E
Word-span	5	9	16	18	19	20				
575-602	3A	1E	2A	2A	2E	2E			4x2	2x2A 2x2E
Word-span	5	9	16	18	19	20				
684-710	1A	1E	2A	2E	2D	2D			4x2	2x2D

is still high in proportion to the number of occurrences of word-spans for individual subjects. Even individual statement descriptors used by subjects have a proportionally high co-incidence, but as Figure 23 shows, when compared with Figure 20 in the previous experiment, the correlations for codes are much less significant in this experiment. Comparisons between the two sets of results are difficult and do not appear to offer any useful basis for statistical analysis. For instance, the word-span 119-219 was identified by eight subjects in this experiment and thirteen in the previous experiment. Both groups produced different correlations. In this experiment five out of the eight subjects thought that this was a ⟨phase one⟩ statement, but were divided on the issue of semantic descriptor - two classified it as ⟨assumption⟩, two as ⟨observation⟩, and one as ⟨hypothesis⟩. In the previous experiment, seven out of thirteen subjects classified it as ⟨hypothesis⟩ within ⟨phase one⟩. Whilst that comparison suggests that ⟨phase one⟩ was popular for both groups, one subject from this group who concurs with seven from the previous group in respect of this single statement classification, does not imply any statistical relevance in the result, which could probably be explained in terms of random selection anyway.

Two aspects of the experiment are important though. First, the subjects have produced non-linear structures, as did the previous group. Second, the subjects, with some degree of discrepency, were able to use the grammar to classify 'chunks' of text. So even if the semantic descriptors themselves are not the most appropriate, the underlying principles of the grammar seem to be. Third, it was thought that a clearer distinction would be seen between results from both groups. In fact, although the non-scientists have not

produced as many statement or classification correlations as the scientists (whom I imagined had more empathy with a conventional format for empirical argument), they have in fact applied the grammar to the sample text in a similar way to the scientists and represented the argument as a non-linear structure as well.

With more flexibility in the grammar and the rules and using a larger group of subjects, a better base for statistical analysis may be possible in the future. A final conclusions from this experiment is that the grammar or something like it, is a useful tool for text analysis to produce the kinds of information structure I am investigating and that the experimental methods used are a valid approach for this research in Information Science.

4.7 SUMMARIES PRODUCED FROM SAMPLE TEXT

The aim of this experiment has been outlined in previous sections but it is really to show two results. First, the subjects are asked to produce a summary of the authors' argument in the sample text. These summaries are then analysed and the individual statements classified using the grammar elements. The sample text has already been shown to have a non-linear argument format, so the structures given in these summaries should be compared with that result. Second, the subjects in this experiment were divided into two groups. The one group (control) had prior knowledge of the grammar, but the other group (experimental) did not. The summaries for these two groups should also be compared to see what structural differences there are between them. For brevity in this Chapter I have produced a table showing the results of the experiment. The whole text of the summaries appear as Appendix E.

The first obvious aspect of the results is that the control group produced structures which were smaller than the experimental group. This could be a consequence of applying the macro structure concept which comes from the grammar, too rigidly. Secondly, although a highly significant result comes from the control group to the effect that linear structures are identified each time, two of the experimental group produced non-linear summaries. This could be a consequence of mis-interpretation of the instructions or a genuine perception of non-linear structures on their behalf. A larger subject group would have supplied more and possibly better data for testing in this experiment, but my conclusion from these results at least, is that VAN DIJK's (1977) theory of structural transformations when humans summarise text from a non-linear to a linear form does have some validity. Once again, the experimental method has been shown, I think, to have a place in this research, as compared with purely theoretical modelling.

FIGURE 24

RESULTS FROM SUMMARIES IN THE CONTROL GROUP

SUBJECT NUMBER	CLASSIFICATION CODES	RESULT
1	1E, 2A, 3A	Linear
2	1E, 2D, 2A, 3A	Linear
3	1E, 2D, 2E, 3A	Linear
4	1A, 2A, 2D	Linear
5	1E, 2A, 2A, 2D, 3A	Linear

FIGURE 25

RESULTS FROM SUMMARIES IN THE EXPERIMENTAL GROUP

SUBJECT NUMBER	CLASSIFICATION CODES	RESULT
1	1A, 1D, 2A, 2B, 2E, 2A, 2D, 3A, 3A	Linear
2	1A, 1D, 1E, 2E, 2D	Linear
3	1E, 1D, 2A, 2D, 2D, 2B, 2A, 2B, 2B, 2D, 3A	Linear
4	1C, 1D, 1D, 2D, 2E, 2D, 1C, 2B, 2A, 3A	Non-linear
5	1E, 2D, 3A, 2A, 2D, 3A, 3A	Non-linear

As can be seen from the tables above, a high proportion of subjects produced linear representations of the authors' argument in the sample text. There can be no comparison of codes used in individual summaries of course, because the text are all dis-similar. The first comparison between the structures in the summaries and the non-linear structure of the whole text, suggests that readers exercise some intellectual transformation of one format into another format, which happens to correspond to the conventional format I am proposing in this thesis. The second comparison between the control and experimental groups shows a high correlation in favour of a linear format from both groups.

4.8 INTERPRETING RESULTS FROM THE EXPERIMENTS

It is now appropriate to recapitulate on the purpose of the experimentation carried out here. The dual aim of the research has first been to determine the existence of an information structure in science text which reflects the organisation of authors' empirical argument and second to propose a model for the process of transferring this information within a writer-to-text-to reader communication system. To determine the first phenomenon entailed creating a set of meta-informational descriptors which could be used to label chunks of text in order to show the organisation being investigated. This set of descriptors has been used in an experimental environment to classify statements in sample text. Although individual subjects obviously identify different chunks of text as being statements, the comparison of descriptors used to classify co-occurring word-spans in the text, suggests that the set of descriptors I have produced is usable and appropriate to demonstrate my thesis that a structure which reflects the organisation of empirical argument in science text exists.

The experimentation has also demonstrated quite clearly that even when section-headings give text some structure, the statements in those sections often do not correspond semantically with the headings themselves. The distribution of statement-types represents what I have called a non-linear format - that is, one which does not follow the linear three-phase progression of my set of descriptors. The three-phase progression represents what I have called the conventional format for argument presentation. To some extent the results given here are an end in themselves, in that they demonstrate the use of the set of descriptors for classifying chunks of text to produce meta-informational structures which reflect the organisation of authors' empirical argument.

There is another aspect of the study for which these results are critical. That is the process by which this meta-information, (or structure reflecting the organisation of the authors 'message'), is transferred to the readers of text. The results shown in the last section (4.7), indicate that readers somehow transform a non-linear structure into a linear one when producing summaries of the initial

text. At this point VAN DIJK's (1977) work makes a significant contribution to my thesis. He has produced what he calls macro-semantic rules. These rules are:

(i) deletion - (so we can disregard superfluous prose);

(ii) generalisation - (so that each statement does not require qualification);

(iii) selection - (so that only statements or words which are appropriate to our overall 'message' can be extracted);

(iv) construction or integration - (so that we can re-order and para-phrase text in order to summarise it).

VAN DIJK says that these four rules are generally used when we summarise a whole text. He gives ample examples of this process which I shall not repeat here but will mention again when I discuss further work in the conclusions of this thesis.

The inference from VAN DIJK's work and my results here is that readers also retain an ideal of the conventional format for argument presentation and that they do in fact carry out transformation of one structure in whole text to another in a summary of the argument contained in the whole text. The following chapter uses this inference to suggest a model for the process of information transfer from writer → text → reader.

4.9 SUMMARY

This chapter has documented the methodology and results of the experimentation I have conducted in this research. I have shown how the assumptions I began with about the nature of empirical argument and its organisation in text, led to a 'pilot study' using a set of semantic descriptors to classify statements thus producing structures which relfect such an organisation. The results from this study showed the worth of producing structures from the analysis of text using this set of descriptors and therefore, I conducted larger experiments to determine more accurately the existence of meta-information in a sample text. The results of these experiments, together with one final experiment where subjects produced written summaries of the

authors' argument in text, suggest that although writers attempt to present their empirical arguments using an ideal format (something like my set of descriptors), in fact they do not achieve this aim because of several influencing factors like the non-linear nature of natural language. Even so, when summarising an author's argument readers tend to invoke some transformational rules which are influenced by the ideal of the conventional format and produce summaries which are more highly structured in terms of this format. This evidence leads us into a theory for the process of information transfer from writer-to-text-to reader where the organisational and communication of empirical argument is concerned.

CHAPTER 5

A THEORY OF INFORMATION TRANSFER

5.1 OVERVIEW OF THIS CHAPTER

Having presented the results of my experiments using the grammar to analyse text in the previous chapter, I shall now outline a theory of the process of information transfer which I believe may occur when readers interpret argument organisations in empirical text. I do not suggest that this theory and the model given in section 5.2 below are based solely on the results produced and presented in the previous chapter. As was pointed out there, the variable nature of the decision-making by candidates in the experiments as to what constitutes a statement or series of statements in a given text and how to label these statements when they are identified, prohibites any result more conclusive than a general overall qualitative inference. In any case, the number of documents analysed here is far too small to detect significant trends in writing style or argument presentation which could substantiate any theory being proposed for general application to questions of scientific communication. Even so, the results do indicate enough about the nature of argument presentation by writers and interpretation by readers to support some speculative theory of the process of information transfer, which I hope helps to illustrate my primary contention about the existence of meta-information and semantic organisational properties in text. The theory which I will shortly be outlining, is also based on the theoretical work of VAN DIJK (1977), SHREIDER (1974) and BELKIN (1977), all mentioned previously and from which some justification is sought here.

In this chapter, I present a model of the information transfer process as I see it. This model is preceded by a diagrammatic

representation of the communication system within which this process resides. After a discussion of the model and theory, I have included in this chapter, the description of a computer system which simulates part of the theory. The computer system carries on a dialogue with users based on the grammar elements given in Chapter 3. The simulation aims to show how a writer would use the grammar when producing a highly structured summary of an empirical argument. The second part of the computer system demonstrates how users could retrieve aspects of authors' arguments from a file of summaries structured by the first program, so that the grammar elements are used as retrieval keys and a search tool.

5.2 A MODEL FOR INFORMATION TRANSFER

We are of course, thinking of the transfer of information from text to readers in this thesis. In such a process there are three major variables to be considered. They are:

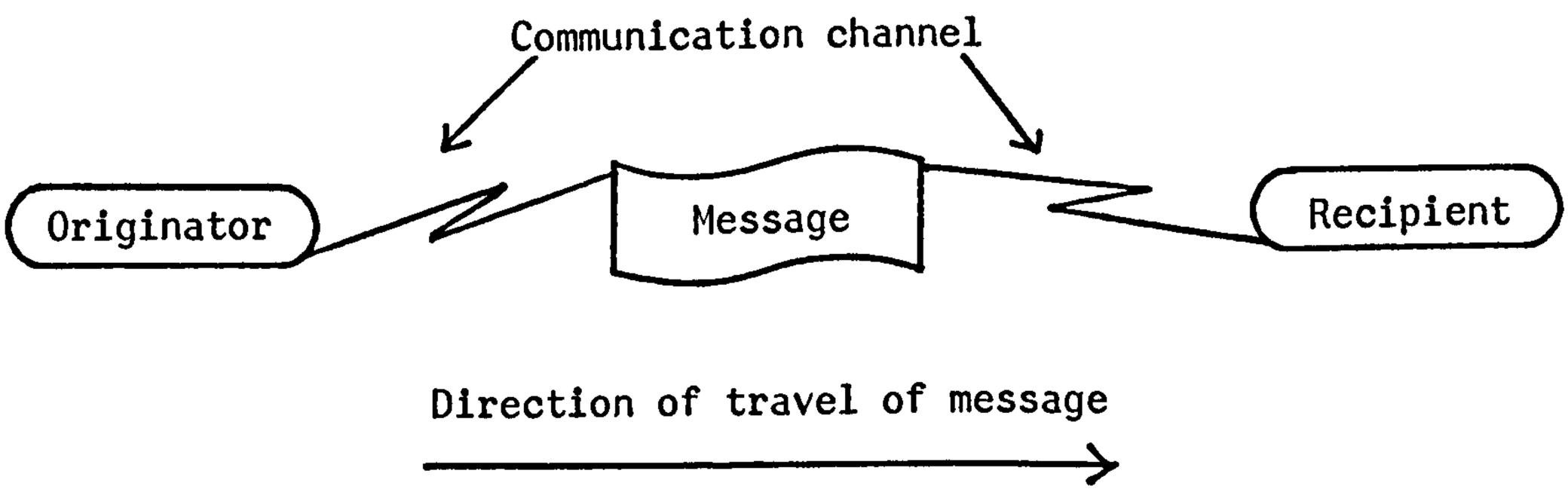
i. the writer or originator of the text (and argument);
ii. the text itself or communication vehicle;
iii. the reader or recipient of the information.

There are, as we shall see, more contributing factors to the communication system we are discussing here, than merely these three variables.

A very simple communication system could be represented by a diagram like the one in Figure 26. In this system, messages are transmitted from the originator to the recipient via some communication channel. The communication channel may be verbal or involve the use of sophisticated hardware to transmit the message. In any case, the message will need to be in a format understandable to the recipient for the system to operate.

FIGURE 26

A MESSAGE TRANSFER MODEL



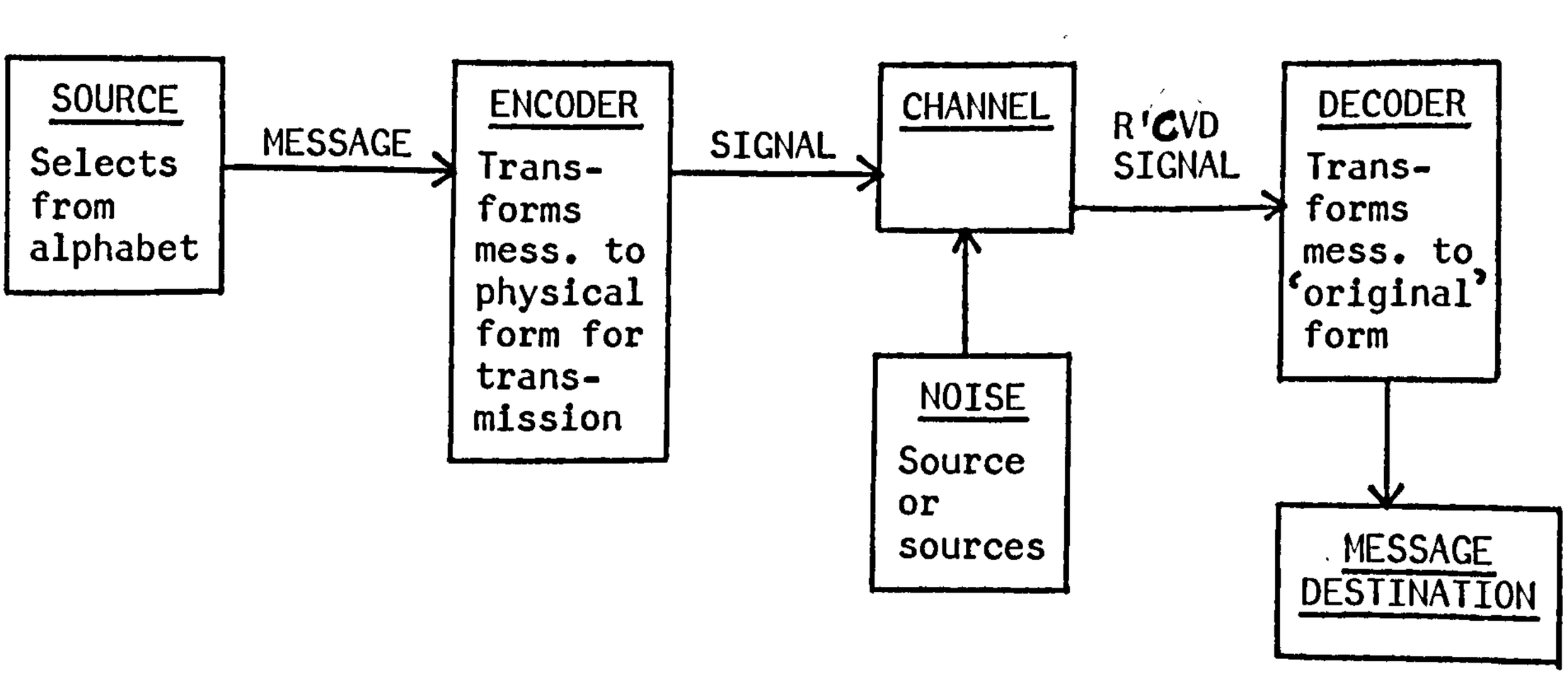
There are two points to be drawn from this example. First, the model does not take into account any input of data or 'ideas' to the originator, or criteria for interpretation of the message by the recipient. Second, there is no indication here of how the message is formulated or presented for communication within the system. If the message was spoken, the format would be constrained by the natural language grammar and vocabulary of the speaker and listener. This would be the case for written natural language too, although the 'ideas' of the originator would have to be set down more formally in one entire presentation because of the absence of interaction and dialogue (questionning and answering) between the originator and the recipient. In the absence of such interaction, written natural language must contain certain semantic 'cues' in order that the message be transferred unambiguously to the reader. As we are well aware, ambiguity is a major problem for both spoken and written natural language. Therefore, if some semantic properties can be easily recognised which reflect the organisation and presentation of the originator's message(s), then it is more likely that the message will be more readily transferred to the recipient. Illustrating this point with a grammar for the presentation of empirical argument in text is the foundation of this thesis. Information itself may be of a conceptual nature, but meta-information is that which reflects the organisation of information in text. As I shall illustrate shortly, information is only latent in text until the text is interpreted, at which point the information is realised. The model in Figure 26 above does not indicate any appreciation of this kind of input to the

communication system. Therefore, neither the format of the message, or that which goes to make up the format by the originator and that which interprets the format by the recipient, is shown in this model. We must build a model which is more sophisticated and contains the properties discussed above.

Much of the nomenclature used by the behavioural sciences when discussing communication systems, has come from SHANNON'S (1948) paper, A mathematical theory of communication. His basic model and the work of WEAVER (1949) which used data from the field of 'Communication Engineering', was concerned with the transmission of electrical pulses. TRAVERS (1970) produced a schematic model of an information transmission process which reflects Shannon's theory and shows most clearly the environment this field of interest refers to. Figure 27 below is a reproduction of Travers' model.

FIGURE 27

AN INFORMATION TRANSMISSION MODEL

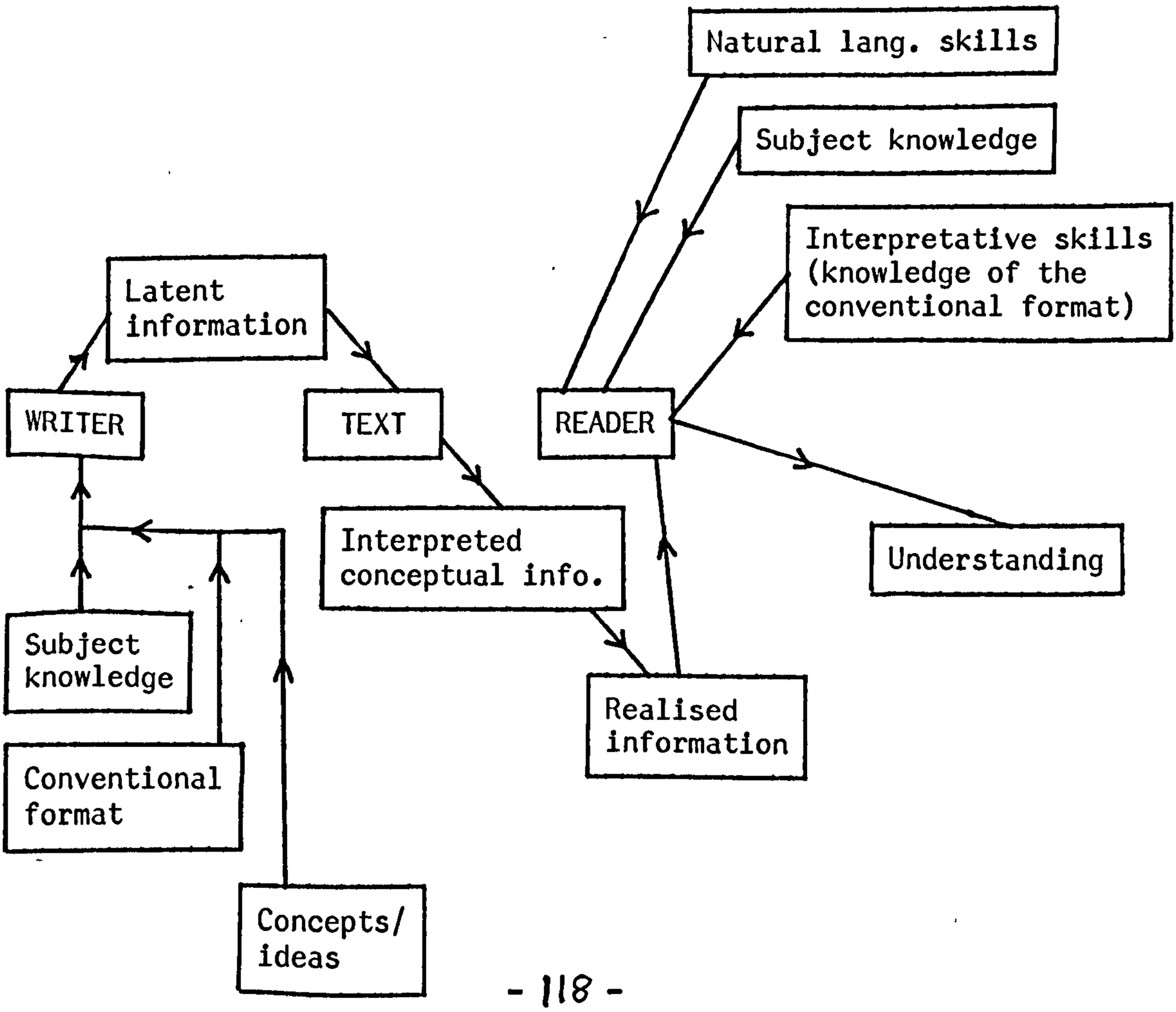


On the whole, this kind of model is concerned with transmitting discrete physical units - usually electrical pulses. Shannon identified variables in the communication process, like sender and receiver, and he also formalised the concept of 'noise' and its effect on the transmission process. This term has come to be used for all kinds of interference in the transmission of signals from sender to receiver. In bibliographic information retrieval, it has

been used to describe elements of information which have 'weak' associations with an element of primary importance which become grouped together with elements of 'stronger' association, thus detracting from the precision of say a search from a data-base for elements of information which have an association with the primary element. In this case, the amount of 'noise' is directly proportional to the number of 'weak' associative elements which have been retrieved with the 'stronger' and more desirable elements. In fact, this model of Travers' goes further than the simple message transfer model which I gave in Figure 26 because it does include external inputs to the primary variables of the system such as sender and receiver. Conceptually it is not flexible enough for me to use directly to illustrate the communication system within which the process of information transfer we are concerned with, resides. In Figure 28 below, I have given a schematic representation of the writer-to text-to reader communication system which operates for this research as I see it. The diagram illustrates how writers messages are formatted and presented in text, then how that format is interpreted by readers of the text. An explanation of the system is given after the diagram.

FIGURE 28

A COMMUNICATION SYSTEM FOR THIS RESEARCH

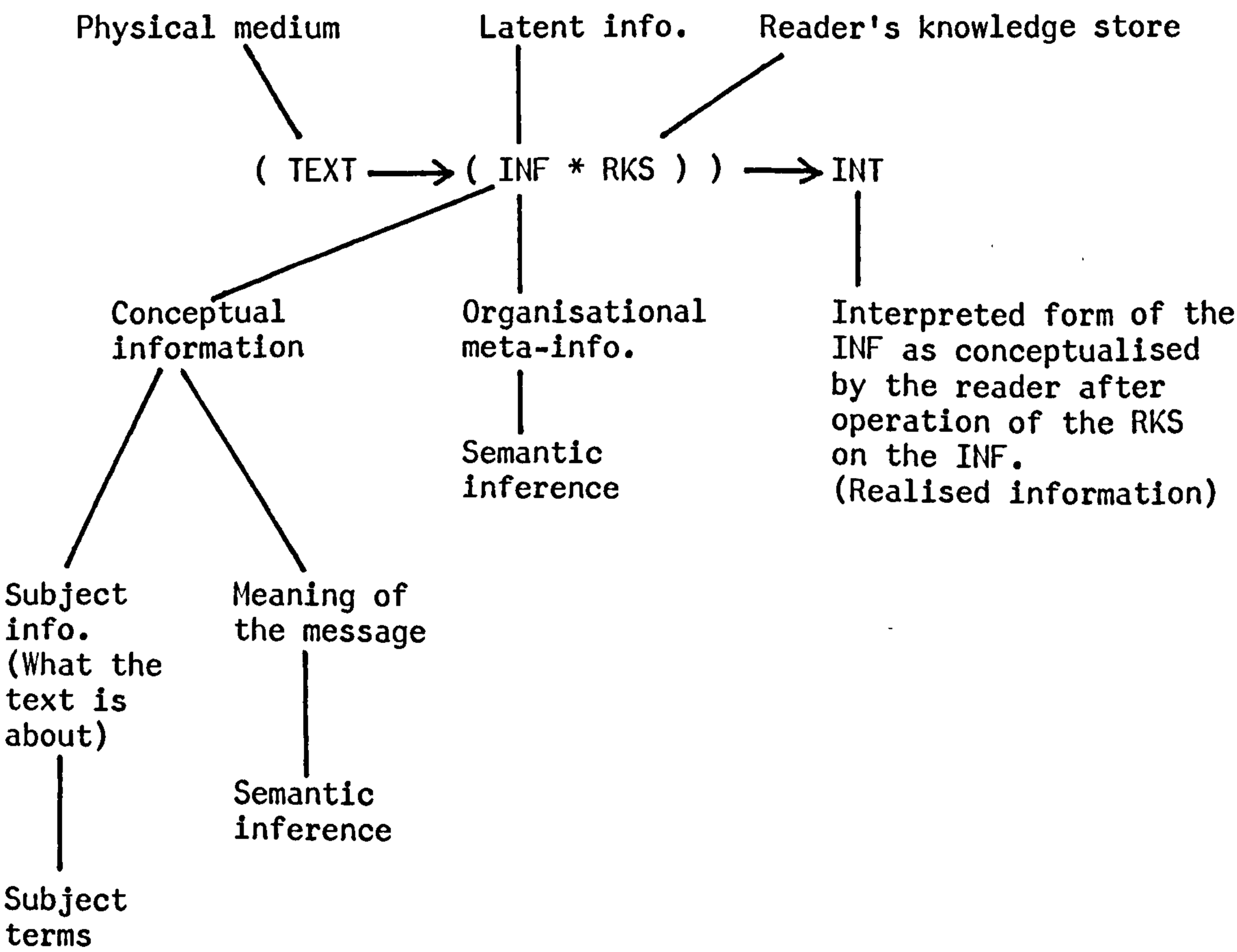


An explanation of this diagram is as follows. Beginning with the writer, I assume that some concept or idea together with subject knowledge and data are formulated to construct an argument. The contents of the argument will form the conceptual information of the eventual written text. Obviously some natural language skills constitute a part of the input to the writer and combined with my assumed ideal of a conventional format, give rise to a written presentation of the argument. All this resides in the physical text as latent information. The reason I differentiate between latent and realised information here, is that I consider the information in text as being latent until the text itself, (and the meta-information), is interpreted by a reader who realises it and synthesises it with his or her knowledge store. On the other side of the physical vehicle, (or Text as shown in the diagram), the reader uses elements of subject knowledge and natural language skills to interpret the conceptual information or 'message' of the writer. The reader's appreciation of a conventional format ideal may assist this interpretation, but will in any case be a tool for interpreting the organisation of the conceptual information. Overall, the interpretation of the text will lead to the latent information being realised and through some process of intellectual synthesis, the message will be understood and perhaps assimilated as new knowledge with existing knowledge.

This, as far as I can see, is the system of communication within which my theory of information transfer resides. I have elaborated on the system to incorporate the notion of a conventional format for empirical arguments; but that aside, I believe the system to be generally applicable to the communication of any information by writers to readers via a natural language medium. It now remains for me to go further and illustrate the process of information transfer which operates within this communication system for the presentation and interpretation of empirical arguments in science text. Figure 29, shows a symbolic model for this process, based partly on the experimental results given in Chapter 4 and to some extent on my own intuitive supposition. The arrows are production or substitution symbols and the asterisk means the operation of one variable on another.

FIGURE 29

SYMBOLIC MODEL OF INFORMATION TRANSFER PROCESS



In the diagram above, I have taken only that part of the communication process which deals with the interpretation of text by readers. In fact, the point I am trying to make here, is that the readers really interpret the information in the text, rather than the text itself. That is, text has associated with it an information structure (see BELKIN (1977) for a theoretical justification of this point), which reflects the content of the author's message. Meta-information, as SHREIDER (1974) describes it, is that part of the information structure which tells us about the conceptual information; or in my case, the organisational properties of the information structure which tells how the message (or conceptual information) is presented. In Figure 29, I show the substitution of the information structure (INF in the diagram) for the TEXT. I have included in the diagram some notes to further illustrate my view of the contents of the INF and other

variables in the model. The information structure (INF) is operated on by the reader's knowledge store (RKS in the diagram). This knowledge store is thought to consist of all those attributes shown as inputs to the READER in the communication system given in Figure 28. That is, subject knowledge, interpretative skills and so on. The operation of the RKS on the INF produces an interpretative structure shown as INT in Figure 29. This interpretative structure is the result of the reader using the RKS to transform a macro-information structure in text (INF), into a micro-information structure (INT), which is a summary of the conceptual information from the INF. I use the term transform here rather than say transpose, because the latter suggests a one-to-one relationship between elements in one structure with elements in the other. It is my view, based on the theoretical work of VAN DIJK (1977) and my own experiments which seem to ratify his results, that a transformational process occurs when writers summarise whole text. Van Dijk showed how four macro-semantic rules could be applied to the summarising process. He calls his rules macro-semantic because they are used to produce a structure which is representative of the whole text or 'whole message' of the author. I mentioned his work in the previous chapter, but to discuss its significance in more depth I have repeated the rules here. They are:

i.	deletion	-	so we can disregard superfluous prose;
ii.	generalisation	-	so that each statement does not individually require qualification;
iii.	selection	-	so that only statements or words which are appropriate to our macro-structure, (main theme of text), can be extracted;
iv.	Construction or integration	-	so that we can re-order and para-phase text in order to summarise it.

Creation could have been used instead of construction to make the intention of the rules more explicit, I think. In fact, although these rules are guidelines for an explanation of what occurs in the summarising process, they are not algorithmic rules. There is no formulation in Van Dijk's paper of algorithms for these rules which could be applied to whole text in order to generate summaries. They are nonetheless useful for getting nearer a definition of how the summarising process occurs and make an important contribution to my theory. What Van Dijk implies by the production of these rules, is that some transformational process occurs and that is precisely what

my experiments show. The change in format from non-linear whole text structure to linear summary must, I feel, be attributable to some rules like Van Dijk's which enable writers to produce results similar to those shown in Chapter Four. Actually, Van Dijk admits that a set of macro-semantic rules is not enough to wholly explain the transformational phenomenon. He mentions (p.139), that at the time of writing he knew of no set of labels for a macro-structure which represent a conventional format in science discourse. Such a set, he assumes, is required to complete the theoretical model. I hope my set of descriptors goes some of the way toward filling this void.

To demonstrate Van Dijk's theory and show how his rules are applied, I have given the following example. In this case, I have taken the first sentence from the summary and the first sentence from the whole text of a recent paper by BROOKES (1978). I have tried to categorise the information from both sentences in such a way that the optimal points of comparison can be seen between the content of each.

Summary sentence:

"The paper outlines a personal view of the cognitive basis of information science."

1. Conceptual information	(Subject info	- "cognitive basis of information science."
	(Functional info.	- "a personal view".
2. Meta-information	(PHASE-ONE ⟨INTRODUCTION⟩)	- "The paper outlines".

Whole text sentence:

"In all practical affairs, communication between humans is mediated by a physical channel."

1. Conceptual information	(Subject info	- "communication between humans is mediated by a physical channel".
	(Functional info	- "In all practical affairs".
2. Meta-information	(PHASE-ONE ⟨ASSUMPTION⟩)	- the whole sentence.

Note that the subject information in the whole text version of the sentence is a statement in its own right. Using Van Dijk's rule of DELETION given earlier, it is likely that the functional information would be disregarded in the summary because an assertive proposition has been made and is most relevant to the argument. The functional information acts only as a qualifier so that the GENERALISATION rule probably applies here as well. I have labelled "a personal view" as functional information in the summary sentence because it appears to me to be a qualifier. It is probably a valid inclusion in summarising the author's message though, because it ensures that readers of the summary will realise this is a personal or intuitive argument rather than the result of say experimental research. I have included my classification of the statements using two different meta-informational descriptors. Note that they are both from PHASE-ONE of the grammar though. Incidentally, when analysed fully, this text, (BROOKES (1978)), produced a linear structure in keeping with the conventional format.

My experiments appear to ratify the previous results of VAN DIJK (1977) as I mentioned earlier. Similarly, his work is some further justification for my own theory of structural transformation in the information transfer process. There is another form of demonstration which shows how this process can be used in practice with the aid of my grammar labels. This is the topic of the next section which outlines a computer simulation of a system to produce highly structured document summaries.

5.3 A COMPUTER SIMULATION OF THE MODEL

Designing a computer system to simulate the model shown in Figure 29 could have led in one of two directions. One way would be to design a system which reads in whole text then processes it to produce summaries using the conventional format. This would either result in a computational experience encountered by others as described in Chapter 2 and shown to produce unsatisfactory results, or I would have to resort to pre-editing the input text which as I have said before proves nothing. In any case such an exercise would go no further in illustrating the use of my grammar for organisational interpretation.

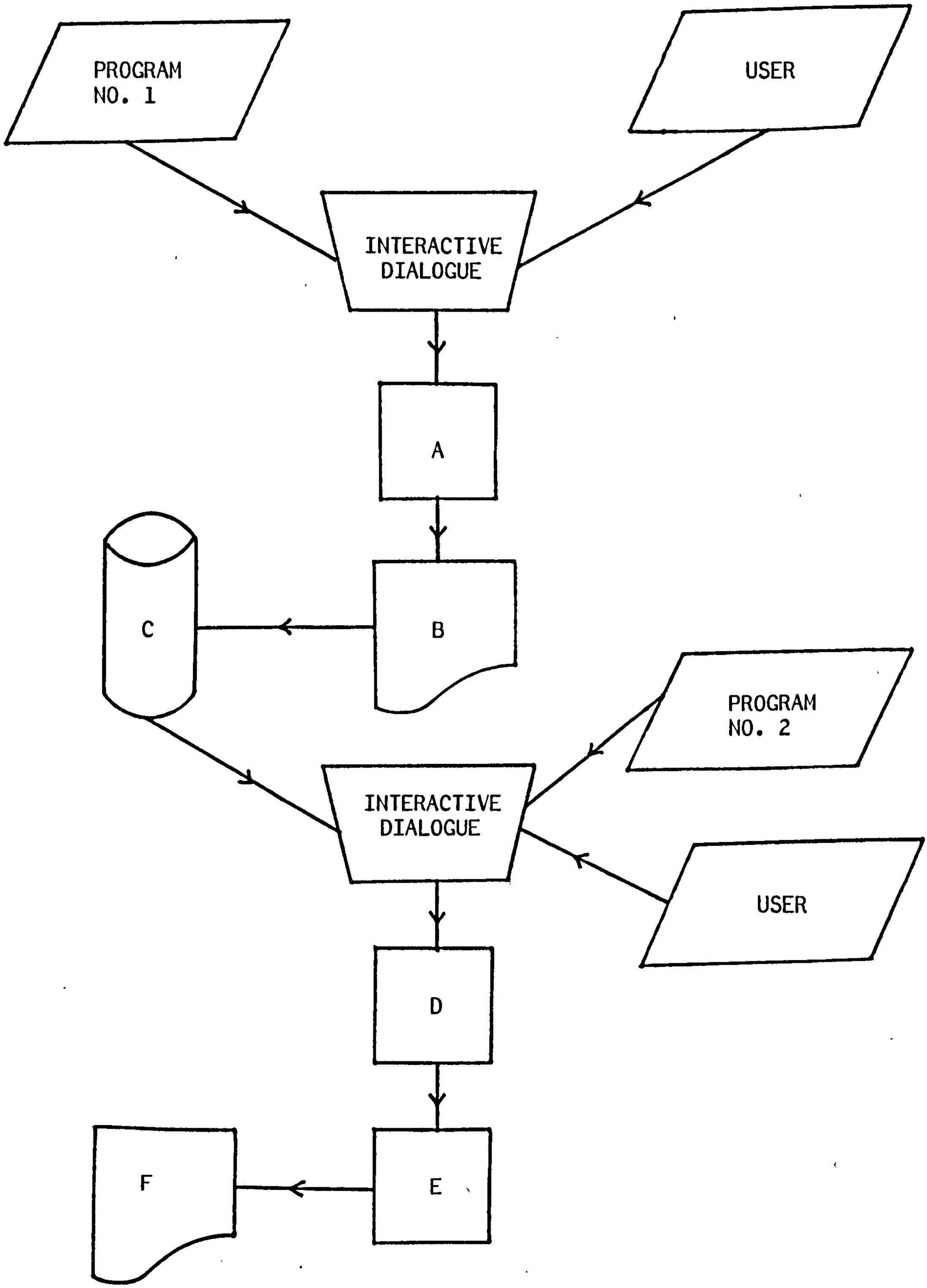
I would merely be attempting to cope with these problems of natural language ambiguity, anaphora and so on, which face researchers into machine translation and the like. The second path to follow is in my view more realistic, if not computationally very elegant.

I have written two programs. The first program carries on an interactive dialogue with a user who, we imagine, wishes to summarise the argument contained in a whole text. The text may or may not have been written by the user. The program puts questions like, "Does your argument have a primary aim?", to the user. In that question we can see present one of the grammar elements - ⟨PRIMARY⟩ . If the user replies in the affirmative to this or any other similar question, the program will ask the user to type in a statement about say, the primary aim of the argument. Having completed the dialogue to the point where questions have been asked relating to all the elements in the form of the grammar, but including only those elements of the grammar which have been used by the user. This program is intended to demonstrate the question and answer situation which I assume occurs, (even if only sub-consciously), when the writer of a summary interprets the information contained in a whole text.

The second program is really an extension of the idea that an interactive computer system which asked questions of users in the way I have just mentioned, might be a useful way of ensuring that documents, (be they empirical argument or technical reports or even service manuals), are produced in a highly structured format. This would be to aid scientific and technical communication, based on some theory of information transfer such as I have given, not to restrict individual writing styles. This second program assumes that we have a file or data-base containing text which has been structured using my grammar elements. The program offers users the choice of any item from the file and allows entry by subject keyword - any form of retrieval access would obviously suffice here. The user is then given the set of grammar elements and asked which aspect of the text is required. The user then inputs a code for the appropriate aspect of the document (or argument), which is immediately output. The program offers a facility for outputting the whole document, selecting another document or listing titles throughout processing. In Figure 30, I have given a diagram showing the flow of work through the system. The

FIGURE 30

SYSTEM FLOWCHART OF THE COMPUTER SIMULATION



system was written in BASIC on a general-purpose micro-computer and the programs are listed in Appendix F.

I have omitted the terminal symbols to start and stop the flow-chart in Figure 30 because I am trying to avoid the diagram from being taken too literally. That is to say, the flowchart has been included to illustrate the inputs, outputs and main processing in the simulation, but as a system the process is not continuous. There are two programs which run indpendently of one another in reality and this flowchart is given to show the overall simulation concept rather than the actual operation of either program. I have used the letters A, B, C, D, E and F to indicate aspects of the overall process. These letters mean:

A - the program asks the user questions about the contents of the argument being summarised. The user inputs data in answer to these questions.

B - when the dialogue has finished, the program outputs a summary using the input from the user together with appropriate grammar elements which are stored in the program.

C - completed summaries are stored on a data-base or file of summaries, indexed by subject keywords.

D - the second program asks the user to input a subject key-word in order to retrieve a summary from the data-base or file of stored summaries.

E - when a summary has been retrieved, the program presents the user with the list of grammar elements and asks the user to use one of them to select some aspect of the author's argument from the summary.

F - output may be either single aspects of the argument or the whole summary as it is stored on the file.

In order to retain some continuity with my examples, I have used the Moon Illusion text which was the subject of analysis in the experiments, as the basis for demonstrating the first of the two computer programs. I have also used one of the individuals who took part in the experiment for writing summaries of the Moon Illusion text, to carry on a dialogue with the program for the example output given in Figure 31. The lines which are preceded by arrows in the left-hand margin are those input by the user. Any line not preceded by an arrow has been generated by the program. The summary which is generated by the program appears at the end of the listing. As can be seen, the program merely concatenates individual grammar elements with the input given by the user. Although some changes in the program would make output summaries even more readable than the one which appears in Figure 31, this listing is sufficient to show the kind of structure which emerges from the summary. Immediately following the output from this program, Figure 32 shows a listing from the dialogue between the second program and a user. If followed through in a similar way to the listing in Figure 31, the retrieval potential and kinds of output given are quite obvious. By virtue of this being a simulation and not a fully implemented system, text for the four titles in the second program has not been stored on the file. Output statements and entire text are simulated by a replacement statement to that effect being output by the program.

What these two programs demonstrate, apart from a inadequate keyboard skills of the users, is a system for constructing summaries of arguments in a format which has some theoretical justification for use as shown by the experiments in Chapter 4. The first program is intended to simulate the activities of individuals as proposed by the information transfer model given earlier in the chapter. The second program demonstrates what could be a potential use for the grammar elements as a retrieval tool if items in a file were structured in such a way that the grammar elements became retrieval keys leading directly to aspects of information in the items. Overall, both of these programs illustrate how meta-information can be seen as distinct from conceptual information and how my grammar elements are an example of this notion.

FIGURE 31

UNEDITED OUTPUT FROM SUMMARY GENERATION PROGRAM

```
•RUN

THIS PROGRAM GENERATES ABSTRACTS FROM DATA SUPPLIED
BY THE AUTHORS OF DOCUMENTS.  QUESTIONS ARE ASKED BY
THE PROGRAM, WHICH THEN GENERATES AN ABSTRACT BASED
ON THAT DATA AND A GRAMMAR WHICH IS HELD IN THE SYSTEM.

DOES YOUR ARTICLE HAVE PRINCIPAL HYPOTHESIS
ANSWER 'Y' OR 'N':

•N

DOES YOUR ARTICLE HAVE PRIMARY AIM
ANSWER 'Y' OR 'N':

•N

DOES YOUR ARTICLE HAVE INTRODUCTORY ASSUMPTION
ANSWER 'Y' OR 'N':

•Y
ENTER THE INTRODUCTORY ASSUMPTION IN NO MORE THAN ONE LINE:

•ILLUSION OF SIZE OF MOON

DOES YOUR ARTICLE HAVE FACT/DATA
ANSWER 'Y' OR 'N':

•APPARENT SIZE INVARIATE WITH VIEWING ANGLE
ENTER THE FACT/DATA IN NO MORE THAN ONE LINE:

•APPARENT SIZE CONSTANT WITH VIEWIND ANGLE

DOES YOUR ARTICLE HAVE CITATION
ANSWER 'Y' OR 'N':

•N

DOES YOUR ARTICLE HAVE METHOD
ANSWER 'Y' OR 'N':

•Y
ENTER THE METHOD IN NO MORE THAN ONE LINE:

•ELIMINATE ATMOSPHERIC DISTORTION BY COMPARING MOON PHOTOGRAPHS

DOES YOUR ARTICLE HAVE RESULTS
ANSWER 'Y' OR 'N':

•Y
ENTER THE RESULTS IN NO MORE THAN ONE LINE:

•FURHER THE HORIZON THE LARGER THE MOON APPEARS

DOES YOUR ARTICLE HAVE METHODOLOGICAL ASSUMPTION
ANSWER 'Y' OR 'N':

•N

DOES YOUR ARTICLE HAVE PRIMARY CONCLUSION
ANSWER 'Y' OR 'N':

•APPARENT SIZE JUDGED AGAINSTSIZE OF FAMILIAR OBJECTS
ENTER THE PRIMARY CONCLUSION IN NO MORE THAN ONE LINE:

•APPARENT SIZE IS JUDGED AGAIST SIZE OF FAMILIAR OBJECTS

DOES YOUR ARTICLE HAVE DEDUCTIVE CONCLUSION
ANSWER 'Y' OR 'N':

•N

DOES YOUR ARTICLE HAVE INDUCTIVE CONCLUSION
ANSWER 'Y' OR 'N':

•N


##### ABSTRACT FOLLOWS:


INTRODUCTORY ASSUMPTION ILLUSION OF SIZE OF MOON.
FACT/DATA APPARENT SIZE CONSTANT WITH VIEWIND ANGLE.
METHOD ELIMINATE ATMOSPHERIC DISTORTION BY COMPARING MOON PHOTOGRAPHS.
RESULTS FURHER THE HORIZON THE LARGER THE MOON APPEARS.
PRIMARY CONCLUSION APPARENT SIZE IS JUDGED AGAIST SIZE OF FAMILIAR OBJEC
TS.
```

FIGURE 32

UNEDITED OUTPUT FROM RETRIEVAL PROGRAM.

```
•RUN

WELCOME TO THE IRS SIMULATION - PLEASE ENTER THE
TERM OF YOUR CHOICE FROM THE FOLLOWING LIST:

COMPILER          MARBLES          PIGS          VEHICLES

•COMPILER

REF: 1/1 - 'COMPILER FAULT TESTING'

DO YOU WANT TO INSPECT THIS DOCUMENT?  TYPE 'Y' OR 'N':

•Y

YOU HAVE INDICATED THAT YOU WOULD LIKE TO INSPECT
THE DOCUMENT CONTAINING THE KEYWORD 'COMPILER'
MORE CLOSELY.   HERE IS A GRAMMAR THAT YOU CAN USE TO
ASCERTAIN THE AUTHOR'S ARGUMENT IN THE DOCUMENT.

1    = HYPOTHESIS
1A   = PRIMARY AIM
1B   = INTRODUCTORY ASSUMPTION
2    = FACT/DATA
2A   = CITATION
2B   = METHOD
2C   = RESULTS
2D   = METHODOLOGICAL ASSUMPTION
3    = CONCLUSION
3A   = DEDUCTIVE CONCLUSION
3B   = INDUCTIVE CONCLUSION

ENTER THE CODE NUMBER WHICH CORRESPONDS TO THE TYPE
OF STATEMENT UYOU WISH TO DISPLAY FROM THE DOCUMENT

•1

THE PRINCIPAL HYPOTHESIS OF THIS TEXT IS:

THIS IS A SAMPLE STATEMENT LINE FROM A TEXT


TYPE '1' FOR ANOTHER GRAMMAR ELEMENT, '2' FOR ANOTHER
TEXT, '3' TO PRINT OUT THE ENTIRE TEXT, OR '4' TO END
THE PROGRAM:

•1A
  ?
INVALID ITEM IN READ OR INPUT
•1

ENTER THE CODE NUMBER WHICH CORRESPONDS TO THE TYPE
OF STATEMENT UYOU WISH TO DISPLAY FROM THE DOCUMENT

•1A
```

FIGURE 32 CONTINUED

```
THE PRIMARY AIM OF THIS TEXT IS:

THIS IS A SAMPLE STATEMENT LINE FROM A TEXT

TYPE '1' FOR ANOTHER GRAMMAR ELEMENT, '2' FOR ANOTHER
TEXT, '3' TO PRINT OUT THE ENTIRE TEXT, OR '4' TO END
THE PROGRAM:

-3


THIS IS A SAMPLE TEXT CONTAINING THE WORD 'COMPILER'

TYPE '1' TO END THE PROGRAM OR '2' IF YOU WISH TO CONTINUE:

-2
WELCOME TO THE IRS SIMULATION - PLEASE ENTER THE
TERM OF YOUR CHOICE FROM THE FOLLOWING LIST:

COMPILER        MARBLES         PIGS            VEHICLES

-PIGS

REF: 1/3 - 'THE FEEDING OF PIGS AND OTHER SUCH FUN'

DO YOU WANT TO INSPECT THIS DOCUMENT?  TYPE 'Y' OR 'N':

-Y

YOU HAVE INDICATED THAT YOU WOULD LIKE TO INSPECT
THE DOCUMENT CONTAINING THE KEYWORD 'PIGS'
MORE CLOSELY.  HERE IS A GRAMMAR THAT YOU CAN USE TO
ASCERTAIN THE AUTHOR'S ARGUMENT IN THE DOCUMENT.

1    = HYPOTHESIS
1A   = PRIMARY AIM
1B   = INTRODUCTORY ASSUMPTION
2    = FACT/DATA
2A   = CITATION
2B   = METHOD
2C   = RESULTS
2D   = METHODOLOGICAL ASSUMPTION
3    = CONCLUSION
3A   = DEDUCTIVE CONCLUSION
3B   = INDUCTIVE CONCLUSION

ENTER THE CODE NUMBER WHICH CORRESPONDS TO THE TYPE
OF STATEMENT YOU WISH TO DISPLAY FROM THE DOCUMENT

-1

THE PRINCIPAL HYPOTHESIS OF THIS TEXT IS:

THIS IS A SAMPLE STATEMENT LINE FROM A TEXT


TYPE '1' FOR ANOTHER GRAMMAR ELEMENT, '2' FOR ANOTHER
TEXT, '3' TO PRINT OUT THE ENTIRE TEXT, OR '4' TO END
THE PROGRAM:

-4


THIS IS THE EMND OF THE IRS SIMULATION PROGRAM.
THANK YOU FOR SHOPPING WITH US TODAY.
```

5.4 SUMMARY

This chapter began by referring to the experimental results given in Chapter 4 and saying how they, together with some theoretical justifications based on the previous work of VAN DIJK (1977), SCHREIDER (1974) and BELKIN (1977), were brought together with my earlier suppositions and speculations to form a theory for the process of information transfer from writer- to text- to reader which I have been investigating. Some theories and models of communication systems were given and a communication system for this research was outlined. Within this communication system I showed the process of information transfer working and gave a symbolic model of the variables and events of the model as I consider them to be. Finally, I outlined a computer system to simulate the model and suggested how my grammar elements might be used as a retrieval tool in a system where highly structured documents were stored.

CHAPTER 6

CONCLUSIONS

6.1 OVERVIEW OF THIS CHAPTER

I do not intend to review each section of the thesis in this chapter - individual summaries for each of the previous five chapters already do that. Instead, I would like to achieve in two endeavours here. First, I want to bring together the major elements of the thesis such as my initial assumptions and the experimental work, to try to show a development in thinking which culminates in the theory of information transfer described in the previous chapter. Second, I would like to take up some of the points I mentioned in the previous chapters, but which were put aside at the time in order that they should not interfere with the main theme of the thesis. I will discuss these points here in the form of further work which could be done using this project as a base. In this context I will also mention some practical uses for the theory I have proposed.

6.2 DEVELOPMENT OF THE THEORY

I began my discussion of research problems for Information Science with a statement that many of our problems appeared to be two-faceted, in that they had some physical connection with hard data but also some abstract properties like meaning, understanding and readers' information needs. My project and the ensuing theory of information transfer seems to typify that phenomenum. Whereas the data are physical objects, (words on paper), they are nonetheless symbolic of some message and information; which itself is an abstract concept. The organisation of that information and the suggestion that we can identify properties in text which reflect that

organisation, is an even more abstract concept. I hope that by this stage, my own argument has made this concept clear and acceptable within a definition of the term meta-information.

In effect, I began with an intuitive assumption that written text must have some semantic properties which reflect the way author's messages are organised and presented. I chose empirical argument as an example of the way messages are presented in text, because it seemed to me to use a well-recognised format. I was to discover that if we build a set of descriptors to identify the organisation of empirical argument in science text, that although the 'ideal' of a conventional format seems to exist, a combination of writing style and other natural language constraints shows an absence of the format in text which are analysed using the semantic descriptors. Further experimentation showed that when individuals summarised whole text which did not reflect the conventional format, they nonetheless produced summaries which did reflect the format. Thus, I was able to demonstrate VAN DIJK's (1977) theory which proposed a set of 'macro-semantic' rules for writers of summaries who operate on whole text to transform its message into a condensed form. My contention has been that what enforces rules such as those VAN DIJK proposes, is an intuitive 'ideal' about how empirical argument, (in my example), should be presented in text.

Beyond all this is yet another notion - that of the difference between information and meta-information. The latter term is one highlighted by the work of SHREIDER (1974) and is discussed in previous chapters. My thoughts concerning the existence of some organisational properties in text which could be identified and labelled with a set of semantic descriptors, led me to Schreider's work, because he proposed meta-information as being that which describes information. I consider conceptual information as being that which relates to the author's message in text and meta-information as that which reflects the presentation, format or organisation of the conceptual information. Early on in the project I had accepted BELKIN's (1977) concept of text always having an invariate information structure associated with it. The point here is that the structure itself is not inflexible, but the rules which govern the existence of the structure are invariate. I would see one of these 'rules' as saying that there must always be some meta-informational properties present in text, for an information

structure to exist at all. That is, all information must have some meta-informational properties in order that it may be informative. Readers of text, in my view, interpret the meta-information in an endeavour to understand the conceptual information. In that case, information consists of conceptual and meta-information.

Fundamental to the study of problems in Information Science is an appreciation of the existence and organisation of information in natural language text. After all, written text is a major medium for the communication of information and knowledge. I made the point at the outset of the thesis, that the identification of properties of information, together with the establishment of some universally recognised corpus of nomenclature and integrated methodology, was needed before the study of problems in Information Science could be thought of as anything approaching a 'science'. The experience of this research has been that several major concepts and terms do exist, but many of them are subjective and open to criticism or interpretation. Similarly, experimental methods are ad hoc. I have endeavoured to add to the corpus of nomenclature, at least some further theoretical justification for the terms information and meta-information as they have been used by BELKIN (1977) and SHREIDER (1974) respectively. The research has, I hope, given some guidelines for identifying types of information in text; be it conceptual or organisational, and made clear the distinction between latent and realised information. The methods used to investigate the existence of semantic properties and to propose a theory for information transfer have been successful here, but may not be generally useful for all investigations. To determine the existence of semantic properties in text meant somehow labelling relevant aspects of it. I chose empirical argument as a form of presentation, to use as an example of how conceptual information was organised in text. Having built a set of descriptors which I could use to label aspects of text, I went about designing experiments to get individuals to produce information structures as they were in the sample text. This and the other experiments were an endeavour to show the set of descriptors at work, but most important to demonstrate the existence of semantic properties within the text. Having carried out the experimental exercises, I proposed my theory of the process of information transfer from text to readers. I felt that it was not enough to merely identify some organisational

properties in text, I needed to show how this meta-information was used in the interpretation of the message in text by readers. In short, my method of investigation has been to propose an example of the phenomenon I have attempted to demonstrate, test that example to produce some experimental results, then use those results together with the theoretical justification of previous researchers, to propose my own theory of the information transfer phenomenum. The result may be a speculative theory, but the model itself is founded on data from my experiments and those of the previous researchers just mentioned; particularly Belkin, Shrieder and Van Dijk. I think that in many respects this is a 'final word' on much research of the problems in Information Science. We can say with certainty, (or sometimes statistical probability), what level of recall a user can expect who is searching for titles in a given subject using a particular bibliographic data-base. That kind of information is based on 'hard' data and is essentially an exercise in statistics. The difficult aspect of the same problem is to endeavour to establish the user's information needs in the first place. If the user is concerned with a particular well-defined topic, the problem does not arise. However, for a research scientist who is hoping to retrieve all citations to the matter to be investigated, some further questioning may be required in order to produce an optimal search strategy. Devising methods and systems for carrying out this latter task, comes into the category of subjectiveness and such abstract concepts as reader's information needs. This is where my research lies too. To some extent I have been able to impose an experimental method on a part of the problem in hand, but mostly I have been dealing with an intellectual process which is by nature volatile and extremely difficult to quantify. These kinds of problems cannot be ignored by Information Science; indeed they are, in my view, fundamental to it, because it is if anything a social science.

6.3 FURTHER WORK

Some of what I have referred to in this thesis have arisen as 'side issues' during the investigation. One such matter is that of the desirability of producing highly structured documents using perhaps some kind of computer system. My computer simulation demonstrates

one kind of output that may be obtained from a system designed to produce highly structured summaries of documents, but is not intended as a prototype for a full-scale system to do something similar. To begin with, my simulation does not take account of the international standards for writing abstracts, or any other form of standards other than that based on my set of descriptors for empirical argument presentation in science text. What I hope I have done is to give some theoretical foundation to the idea of producing highly structured summaries of documents in a form which appears to be an intuitive convention. It is possible, I think, to build a system based either on my set of descriptors or some other format labels, to produce either summaries or original documents like service manuals and reports. I intend to treat this possibility as a separate issue from the next I will mention, which is more speculative.

As a next stage in my research after this thesis, I hope to formalise a grammar using rules like those of VAN DIJK (1977), together with a set of format delimitors something like my present set of descriptors, to develop some algorithms for the analysis of whole text by computer which will produce summaries as output. To solve the many linguistic problems involved would require expertise in language analysis such as that referred to in the work of SAGER (1975) and KITTREDGE (1978). Another possibility for my theory, is to use it in association with the work of VAN DIJK (1977) and attempt to formulate some algorithms for a system to read whole text by computer which would then produce summaries of it. As I mentioned earlier in this chapter, the work of SAGER (1975) and KITTREDGE (1978) with sub-language grammars could be useful here. A good deal of work with knowledge representations and other devices of Artificial Intelligence would also be necessary before a system which was conceptually sophisticated enough could be produced. Similar work by several researchers who are attempting to analyse text to represent meaning, knowledge, understanding and information, has been underway for some years now. It shows every sign of continuing. I hope that this thesis has added something to the theoretical foundations of the area in general and Information Science in particular.

The results produced from the experiments described in Chapter Four were not very suitable for statistical analysis. As the chi-square text used for data in Part II of the Pilot Study showed though, some statistical significance can be found in some of the results. More data and different experimental controls may produce better data for analysis within the methodology used here and an endeavour to provide new data will be made in the near future. If the grammar can be implemented satisfactorily, analysis by computer may produce this data without using human subjects. One of the aims of the methodology in this project was to show that humans could use the grammar, however, and without a suitable response from their analyses, we are in danger of creating experiments which are too artifical for general acceptance. A criterion for estimating the value of results is required before this implementation takes place and that, followed by attempts to analyse whole text using the grammar in a computer system, will be the next stage for my research in this problem area.

6.4 SUMMARY

This chapter follows through the development of the project from my initial assumptions concerning the existence of semantic properties in text which reflect the organisation of authors' messages, to the production of a theory for the transfer of information from text to readers. It brings together the theoretical and experimental evidence, and claims to have demonstrated some of the theory I have used to justify my conclusions. The chapter points to my remarks concerning the need for more substantial definitions and nomenclature in Information Science and suggests how this thesis has contributed to that corpus. Finally the chapter looks at the possibilities for use of my theory in two kinds of system to produce summaries of whole documents.

APPENDIX A

A RANDOM SENTENCE GENERATOR

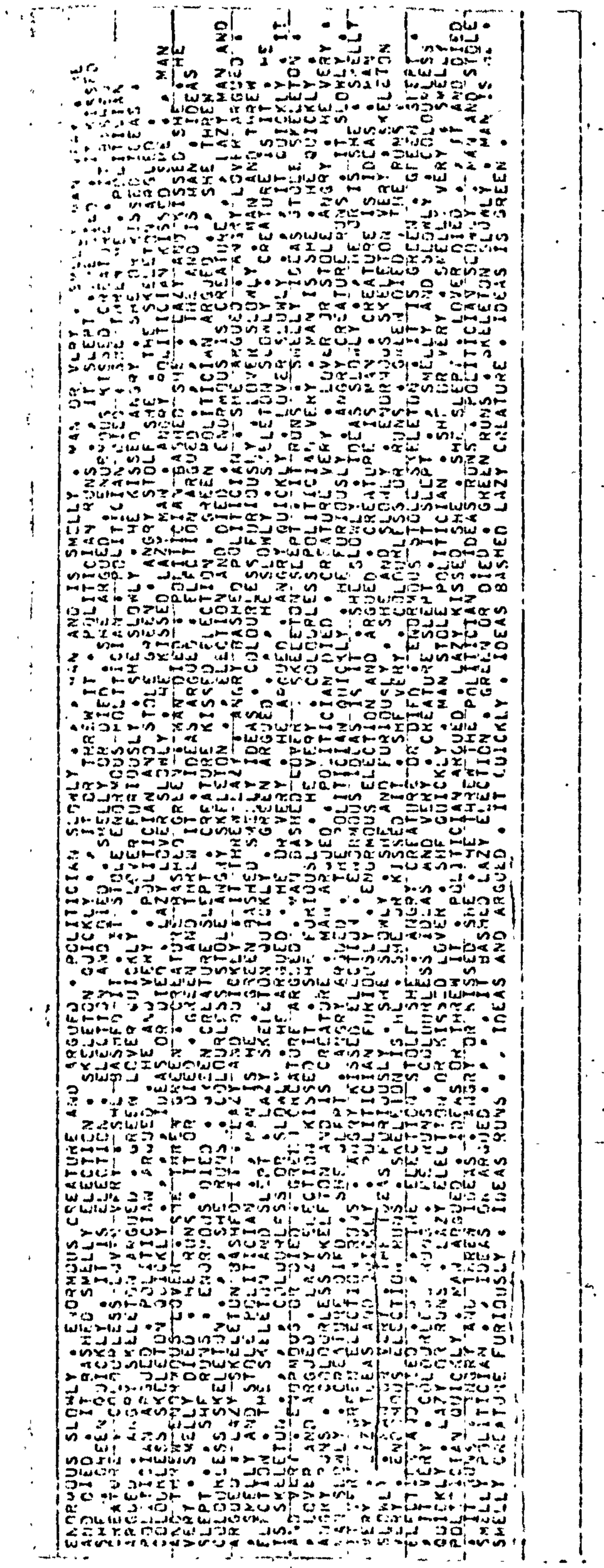
Following is a listing of a computer program written in Algol 60, to demonstrate how a strict grammar representing the formal language structures of the English language (from Chomsky (1957)), can produce semantically ambiguous or even nonsensical sentences when words are selected at random from a valid vocabulary. The algorithm is very simple. A random number (pseudo-random) is generated by the computer from a seed in the program; variable R:= 1234567. The random number is then passed to a set of procedures which in turn randomly select words from internal sets; shown for example as PROCEDURE CONJ (for conjunctions) which has the choice of AND or OR whenever it is called. Procedures are called recursively but are controlled by the value of the random number or when a period is reached. The sentences produced, (output follows the programme listing), are short and often grammatically correct but semantic nonsense. 'HE ARGUED' is fine, whereas 'SHE SLOWLY' is quite wrong. 'HE STOLE ENORMOUS POLITICIAN' is not bad, and 'HE THREW THE POLITICIAN' is good. There is a problem with the random intrinsic, in that the program occasionally seems to generate one or two commas more than are needed.

P H I L I P

```
BEGIN
%  NAME:    GENERATE.
%  TASK:    TO GENERATE ENGLISH SENTENCES AT RANDOM FROM A
%           GIVEN GRAMMAR.
%  METHOD:  USING THE RANDOM INTRINSIC ON THE B6700, THIS
%           PROGRAM CHOOSES ONE WORD AT A TIME FROM THE
%           COMPONENT PARTS OF A SENTENCE (EG: NOUN ), AND
%           COMPILES A SENTENCE UNTIL A PERIOD IS GENERATED.
%           A LIMIT OF  200  SENTENCES HAS BEEN PLACED ON
%           THE GENERATOR.
%  AUTHOR:  PHILIP J. SALLIS, COMPUTER SCIENCE, SPRING 1976.

  FILE CR(KIND=READER), LP(KIND=PRINTER, UNITS = CHARACTERS);

  ARRAY OUTOF[0:132+10];
  POINTER P, Q;
  REAL LL,R;

    DEFINE
      OUTPUT(X) = BEGIN
                    REPLACE WIP BY X," ";
                    IF DELTA(OUTOF,Q) >=132 THEN
                      BEGIN
                        WRITE (LP, LL, OUTOF);
                        P:= OUTOF;
                        LL:= 0;
                        REPLACE WIP BY X," "
                      END;
                        LL:= * + DELTA(P,Q);
                        P:= Q;
                  END#,
      PICK(A) = ENTIER(RANDOM(R) * (A)); #

    PROCEDURE CONJ;
      CASE PICK(2) OF
        BEGIN
          0:  OUTPUT ("AND");
          1:  OUTPUT ("OR");
        END;

    PROCEDURE N;
      CASE PICK(7) OF
        BEGIN
          0:  OUTPUT ("LOVER");
          1:  OUTPUT ("MAN");
          2:  OUTPUT ("POLITICIAN");
          3:  OUTPUT ("ELECTION");
          4:  OUTPUT ("CREATURE");
          5:  OUTPUT ("SKELETON");
          6:  OUTPUT ("IDEAS");
        END;

    PROCEDURE PRO;
      CASE PICK(3) OF
        BEGIN
          0:  OUTPUT ("HE");
          1:  OUTPUT ("SHE");
          2:  OUTPUT ("IT");
        END;

    PROCEDURE ADJ;
      CASE PICK(7) OF
        BEGIN
          0:  OUTPUT ("ENORMOUS");
          1:  OUTPUT ("LAZY");
          2:  OUTPUT ("GREEN");
          3:  OUTPUT ("SMELLY");
          4:  OUTPUT ("THE");
          5:  OUTPUT ("COLOURLESS");
          6:  OUTPUT ("ANGRY");
        END;

    PROCEDURE ADV;
      CASE PICK(4) OF
        BEGIN
          0:  OUTPUT ("QUICKLY");
          1:  OUTPUT ("FURIOUSLY");
          2:  OUTPUT ("VERY");
          3:  OUTPUT ("SLOWLY");
        END;

    PROCEDURE TV;
      CASE PICK(5) OF
        BEGIN
          0:  OUTPUT ("IS");
          1:  OUTPUT ("STOLE");
          2:  OUTPUT ("THROWN");
          3:  OUTPUT ("KISSED");
          4:  OUTPUT ("WASHED");
        END;
```

```
PROCEDURE IT;
    CASE PICK(3) OF
        BEGIN
            0:  OUTPUT ("SLEPT");
            1:  OUTPUT ("RUNS");
            2:  OUTPUT ("DIED");
            3:  OUTPUT ("ARGUED");
        END;

PROCEDURE ON;
    FORWARD;

PROCEDURE NPHR;
    CASE PICK(3) OF
        BEGIN
            0:  N;
            1:  PN;
            2:  ON;
        END;

PROCEDURE PRED;
    CASE PICK(3) OF
        BEGIN
            0:  TV;
                NPHR;
            1:  IT;
            2:  ADV;
        END;

PROCEDURE ADJPHR;
    CASE PICK(3) OF
        BEGIN
            0:  ADJ;
            1:  ADJPHR;
        END;

PROCEDURE ON;
    CASE PICK(3) OF
        BEGIN
            0:  ADJPHR;
                N;
            1:  ADJPHR;
            2:  ON;
        END;

PROCEDURE COMPSUBJ;
    CASE PICK(3) OF
        BEGIN
            0:  NPHR;
                CONJ;
            1:  NPHR;
            2:  OUTPUT (",");
                COMPSUBJ;
        END;

PROCEDURE SUBJ;
    CASE PICK(3) OF
        BEGIN
            0:  NPHR;
            1:  COMPSUBJ;
        END;

PROCEDURE SEN;
    BEGIN
        IF RANDOM(R) < 1/2 THEN
            NPHR
            ELSE
                COMPSUBJ;
                PRED;
                OUTPUT (".");
    END;


%%%%%%%%%% MAINLINE BEGINS %%%%%%%%%%

    LL := 0;                    % FOR "P".
    P := OUTOF;                 % POINT "P" SOMEWHERE.
    R := 12345671;              % LARGE ODD NUMBERS, FOR RANDOM.
    THRU 200 DO
        IF LL = 0 THEN
            WRITE (IP,LL,OUTOF);

%%%%%%%%%% MAINLINE ENDS %%%%%%%%%%

END.

ERRORS DETECTED = 0.
SEGMENTS = 2,  TOTAL SEGMENT SIZE = 1017 WORDS.  CORE ESTIMATE = 2503 WORDS.  STAC
121 CARDS, 2369 SYNTACTIC ITEMS, 46 DISK SEGMENTS.
AMES (ANDROS)PHILIP ON PACK.
ME = 12.902 SECONDS ELAPSED)  5.125 SECONDS PROCESSING)  4.164 SECONDS I/O.
```

OUTPUT FROM THE PROGRAM



APPENDIX B

STATEMENT-TYPE IDENTIFICATION GRAPHS

Statement-type identification within the INTRODUCTORY sections of 20 text

Following are 17 graphs which are given to illustrate the distribution of statement-types in the INTRODUCTORY sections of 20 text. The first three graphs are included in an explanation of this experiment in Chapter 4. The sequence in this appendix begins therefore, at Text No: 4 and ends with Text No: 20. The sequential numbers along the horizontal axis represent individual statements in sequence as they appear in the text. The grammar categories are listed down the vertical axis. Dotted lines across the graph are given to show when the continuous line, which joins the plotted statement-type categories, moves from one phase of argument to another. This represents the non-linear nature of author's arguments in conflict with the conventional (or linear), format of the grammar. The graphs are preceded by an explanation of the significance of the plotted line and statement-type distribution for each individual analysis. I have used 'introductory type' to mean within ⟨ PHASE-ONE ⟩ of the grammar; ⟨ INTRODUCTION ⟩ type means classified with that label. I have done this because the grammar says that labels within phases are considered synomymous for comparisons of linearity. That is, a text may reflect a linear argument even though various labels with individual phases have been used in a 'non-linear' way. Non-linear really means leaving one phase and entering another. I am concerned with comparing statement-type with the section heading of INTRODUCTION here, so all those statements classified within ⟨ PHASE-ONE ⟩ are considered as being of 'introductory type' for the purpose of the analysis.

DISCUSSION OF GRAPHS FOR TEXT NO: 4 TO 25

TEXT NO: 4

Number of statements identified = 12
Number of introductory type = 8

One-third of this argument was classified as being within the second phase of the grammar. Although this section of the text was labelled INTRODUCTORY, only six statements (one-half) out of twelve were classified as being of an introductory nature. This is a non-linear statement-type distribution.

TEXT NO: 5

Number of statements identified = 10
Number of introductory type = 1

Although this section of text was entitled INTRODUCTION, only one statement was classified as being within ⟨PHASE ONE⟩, let alone ⟨INTRODUCTION⟩. Nine out of ten statements were classified as being concerned with methodology and evidence for the author's argument. This is a non-linear statement-type distribution.

TEXT NO: 6

Number of statements identified = 10
Number of introductory type = 4

This text displayed a wide distribution of statement-types ranging from ⟨ASSUMPTION⟩ to ⟨RESULT⟩; 60% of this introductory section referred to methods used, evidence and results, while only one statement was classified as being introductory and three others as assumptions. This is a non-linear statement-type distribution.

TEXT NO: 7

Number of statements identified = 23
Number of introductory type = 2

As the accompanying graph shows, most of this 'introductory' section dealt with evidence for the author's argument and results. Although it is a non-linear statement-type distribution, the graph illustrates an almost linear representation within ⟨PHASE-TWO⟩ of the grammar, leaving out ⟨PHASE-ONE⟩ altogether.

TEXT NO: 8

Number of statements identified = 14
Number of introductory type = 2

This text only reflected statements of introduction after the first six statements, which were concerned with results, evaluation and evidence for the argument. A non-linear representation.

TEXT NO: 9

Number of statements identified = 20
Number of introductory type = 11

This text displays a greater proportion of statements of introduction than of statements within ⟨PHASE-TWO⟩. However, only four of the eleven statements that were classified as being within ⟨PHASE-ONE⟩ were of ⟨INTRODUCTION⟩ type. The same proportion were of ⟨CITATION⟩ type. This is a non-linear distribution.

TEXT NO: 10

Number of statements identified = 6
Number of introductory type = 0

Although only a short piece of text, one would have expected at least one statement within the introductory section to be of ⟨INTRODUCTION⟩ type. Not so here. Difficult to say this is a non-linear distribution because all six statements are within ⟨PHASE-TWO⟩ of the grammar. However, the grammar defines a linear representation as being one which proceeds from ⟨PHASE-ONE⟩ to ⟨PHASE-TWO⟩ to ⟨PHASE-THREE⟩ which this distribution does not do. Furthermore, we are mostly concerned here with comparing the section-heading of INTRODUCTION with ⟨PHASE-ONE⟩ or more particularly ⟨INTRODUCTION⟩ type statements. In this case, none of the statements were classified as being either of those two categories.

TEXT NO: 11

Number of statements identified = 15
Number of introductory type = 2

Most of this text, (nine statements), is concerned with statements of ⟨METHOD⟩ and ⟨RESULT⟩. Only one statement is of an introductory type, with one other in ⟨PHASE-ONE⟩ as an ⟨ASSUMPTION⟩. This is a non-linear representation.

TEXT NO: 12

Number of statements identified = 5
Number of introductory type = 5

The shortest text in this sample, the statements in this section are all within ⟨PHASE-ONE⟩ of the grammar - although none are actually of ⟨INTRODUCTION⟩ type. The author makes three assumptions, one observation then states an hypothesis. This is the kind of linearity suggested by the grammar as being 'ideal'.

TEXT NO: 13

Number of statements identified = 21
Number of introductory type = 10

This argument began with introductory-type statements then left ⟨PHASE-ONE⟩ to make statements of evidence, method and result. The last four statements were actually ⟨INTRODUCTION⟩ type which infers that the author really was trying to introduce the argument and give an overview at the same time. A non-linear representation.

TEXT NO: 14

Number of statements identified = 25
Number of introductory type = 6

Less than twenty-five percent of the statements in this section of text were classified within ⟨PHASE-ONE⟩. Five out of six of them were ⟨INTRODUCTION⟩ type though ⟨METHOD⟩ and ⟨RESULT⟩ dominate the classification within ⟨PHASE-TWO⟩. A non-linear representation. This argument did begin well with four statements of introduction followed by one of assumption. If the author had then begun a second section entitled 'method' or some such label, the next thirteen statements would have reflected a near 'ideal' format in terms of the grammar.

TEXT NO: 15

Number of statements identified = 8
Number of introductory type = 3

This argument began with an ⟨INTRODUCTION⟩ and finished the section, (entitled introduction), with an ⟨AIM⟩. The section was however, non-linear because the author included statements of evidence, method and result, between these two introductory-type statements.

TEXT NO: 16

Number of statements identified = 11
Number of introductory type = 7

Six of the seven ⟨PHASE-ONE⟩ statements were of ⟨ASSUMPTION⟩ type. The seventh (and last in the section), was of ⟨INTRODUCTION⟩ type. This in itself would be well within the 'ideal' of the grammar, but the author interspersed these statements with ones of evidence and result. Although these latter statements were to support the assumptions made, the author could have stated the assumptions in the introductory section, then referred back to them in the second phase of the argument where evidence and data for it are most appropriate.

TEXT NO: 17

Number of statements identified = 11
Number of introductory type = 5

This text begins in ⟨PHASE-TWO⟩ within a statement of ⟨EVIDENCE⟩ and more than fifty percent of the statements remain in that phase. Two of the five ⟨PHASE-ONE⟩ statements are ⟨INTRODUCTION⟩ type, but even the last statement is ⟨RESULT⟩ type. This distribution is not only non-linear, but also well away from the 'ideal' format of the grammar.

TEXT NO: 18

Number of statements identified = 8
Number of introductory type = 4

Only one of the four statements classified within ⟨PHASE-ONE⟩, was of ⟨INTRODUCTION⟩ type and that second to last in the text. There is an ⟨AIM⟩ in this argument, but it occurs after statements of evidence and result. A non-linear distribution, but it did begin in ⟨PHASE-ONE⟩ with a statement of ⟨ASSUMPTION⟩.

TEXT NO: 19

Number of statements identified = 16
Number of introductory type = 8

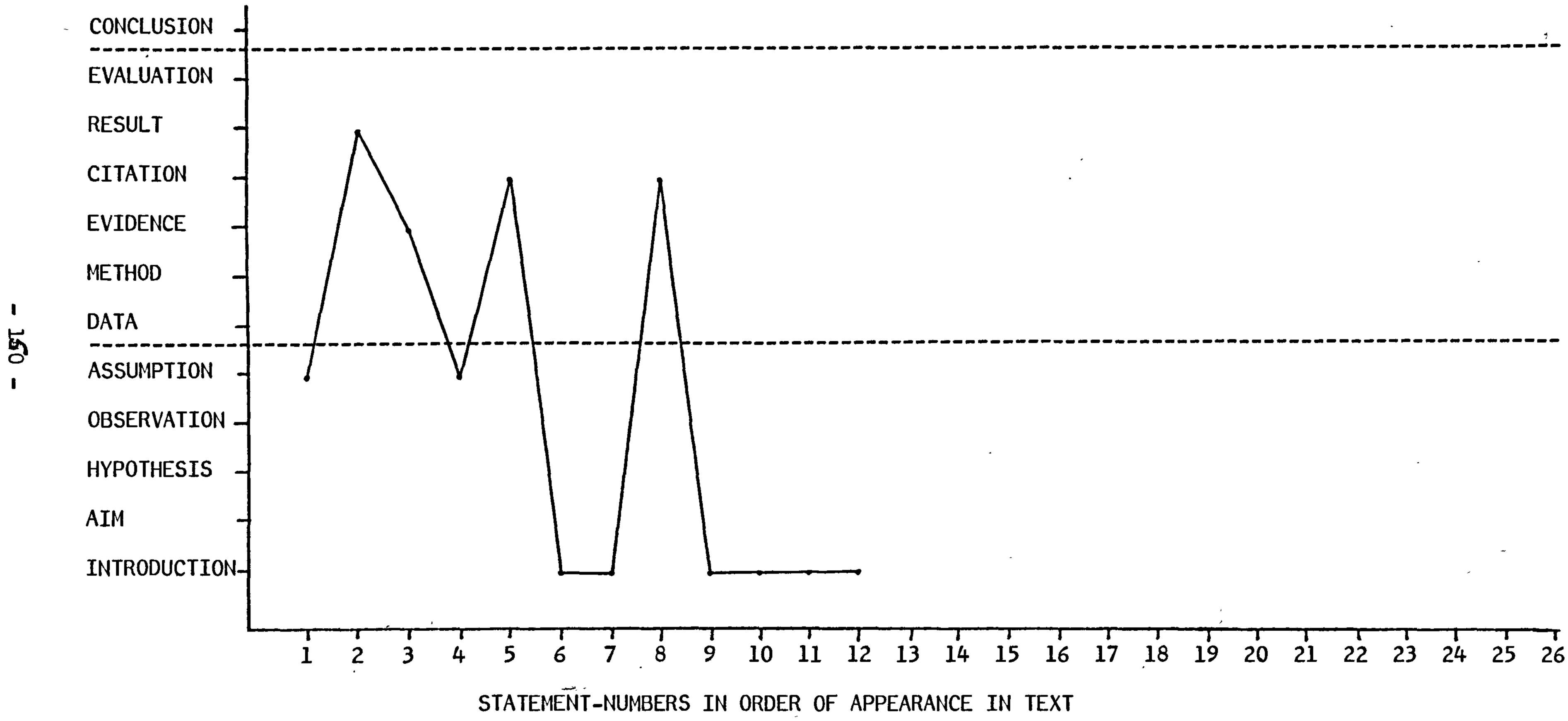
One-half of the statement in this text were of introductory type. When reading it, the section could have been divided into ⟨INTRODUCTION⟩ and ⟨METHOD⟩ after the ninth statement but even then two statements would have been outside of the format and one introductory type occurred in the last group of three statements.

TEXT NO: 20

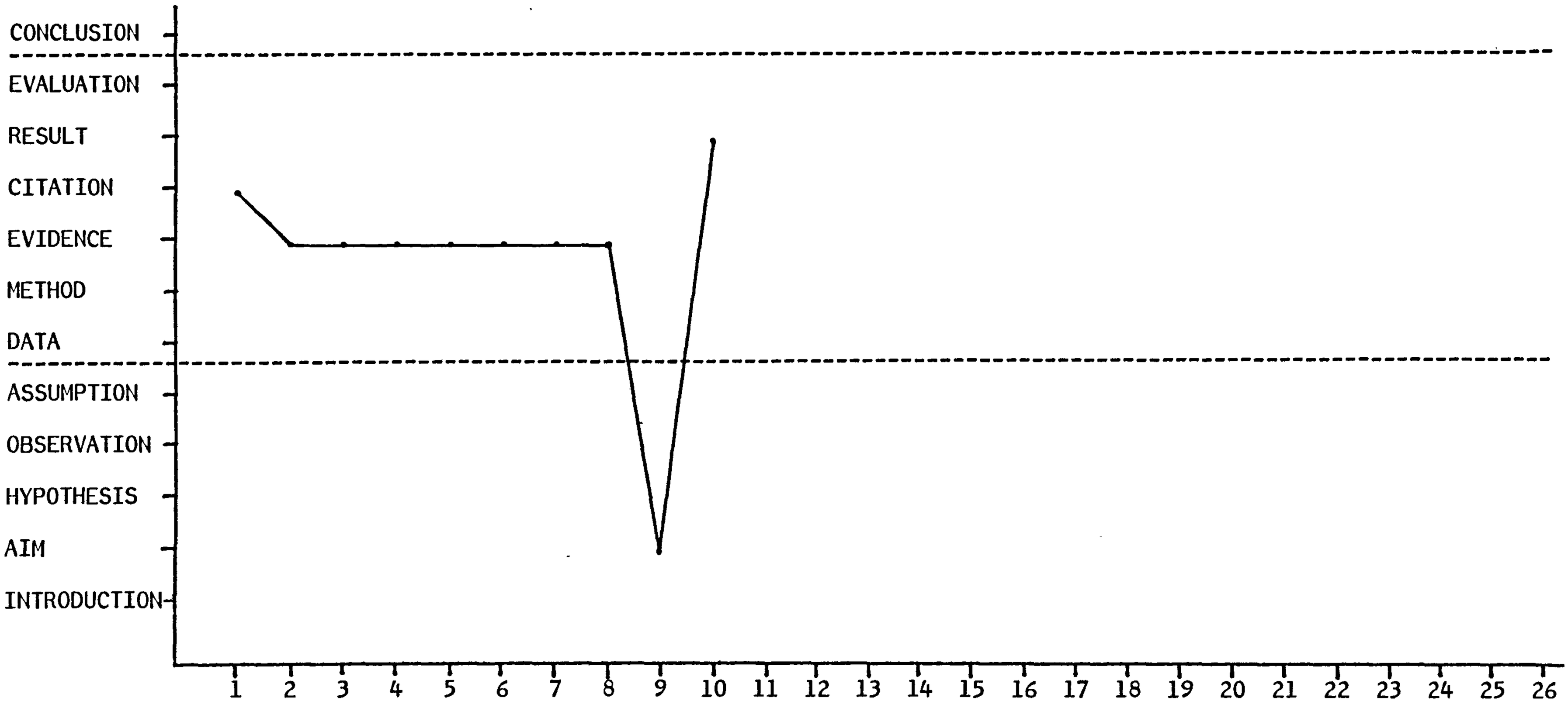
Number of statements identified = 12
Number of introductory type = 5

Most of this text dealt with previous work but nevertheless, the statements were mostly of ⟨PHASE-TWO⟩ type. This accounts for six of the statements being of ⟨RESULT⟩ and ⟨EVALUATION⟩ type.

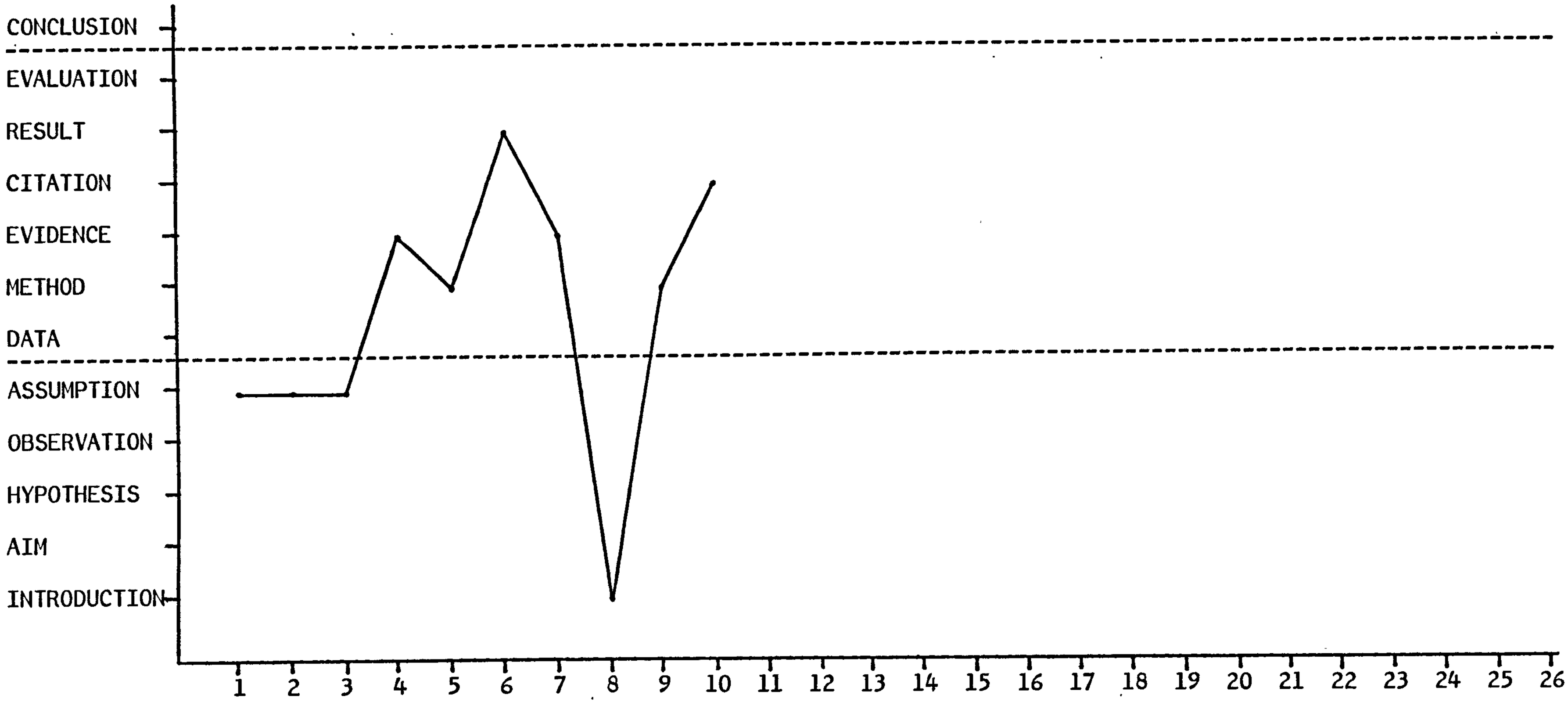
TEXT NO: 4


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

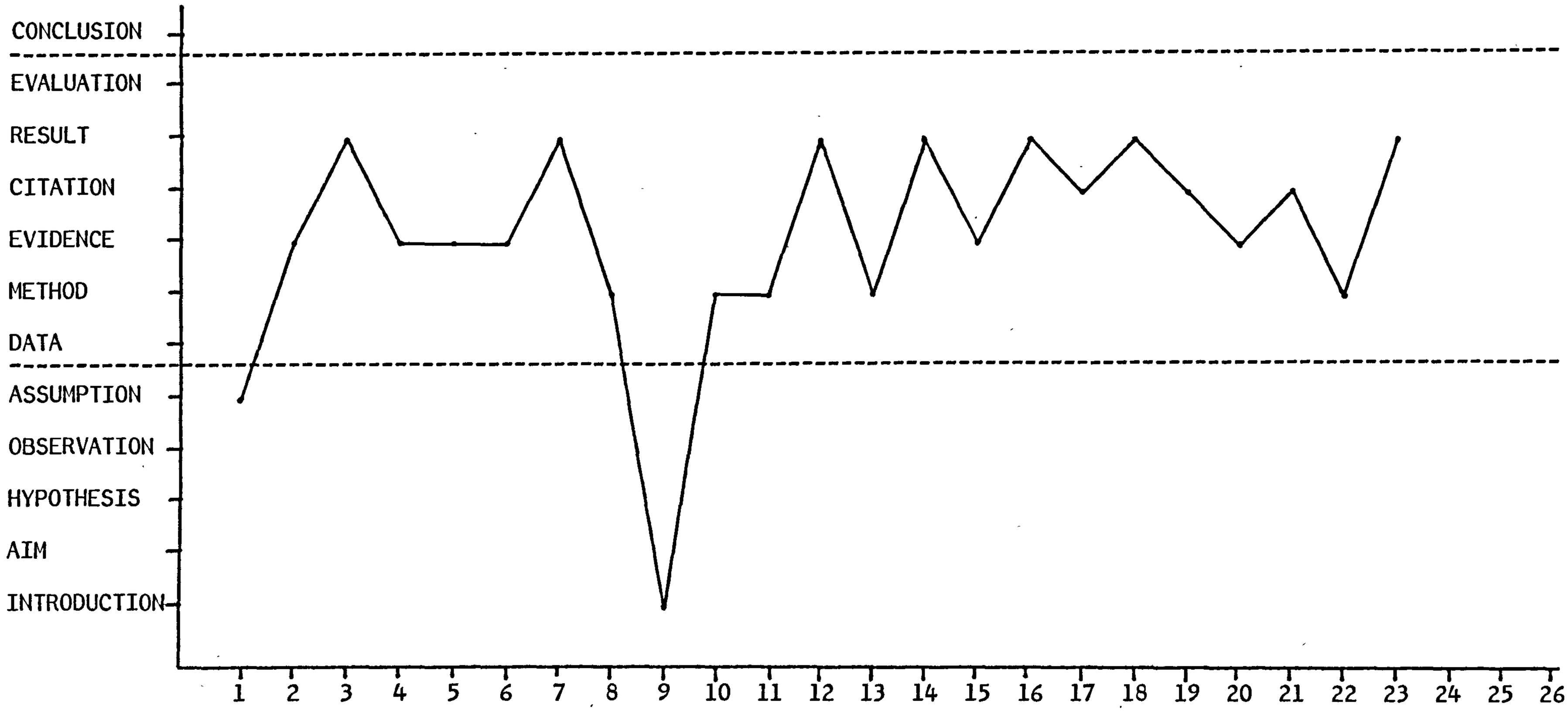
TEXT NO: 5


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

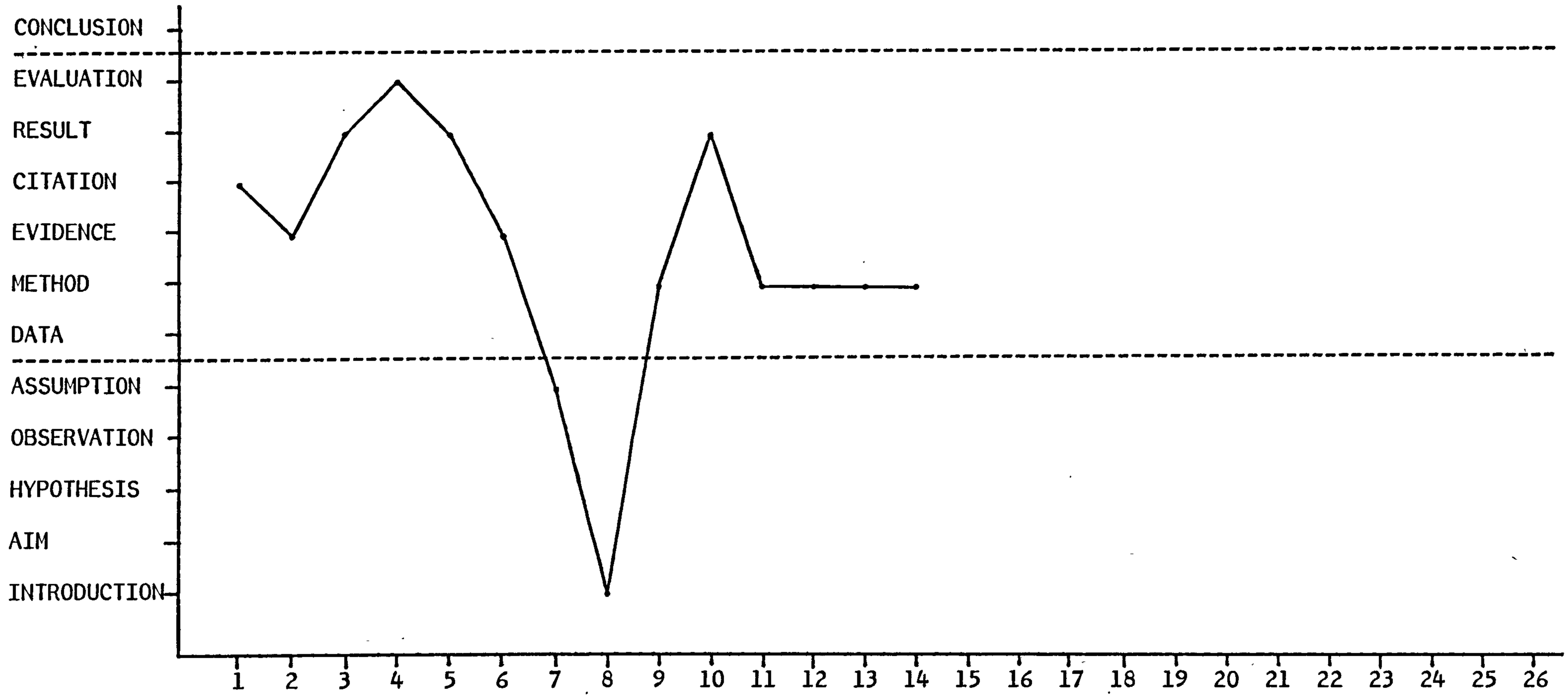
TEXT NO: 6


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

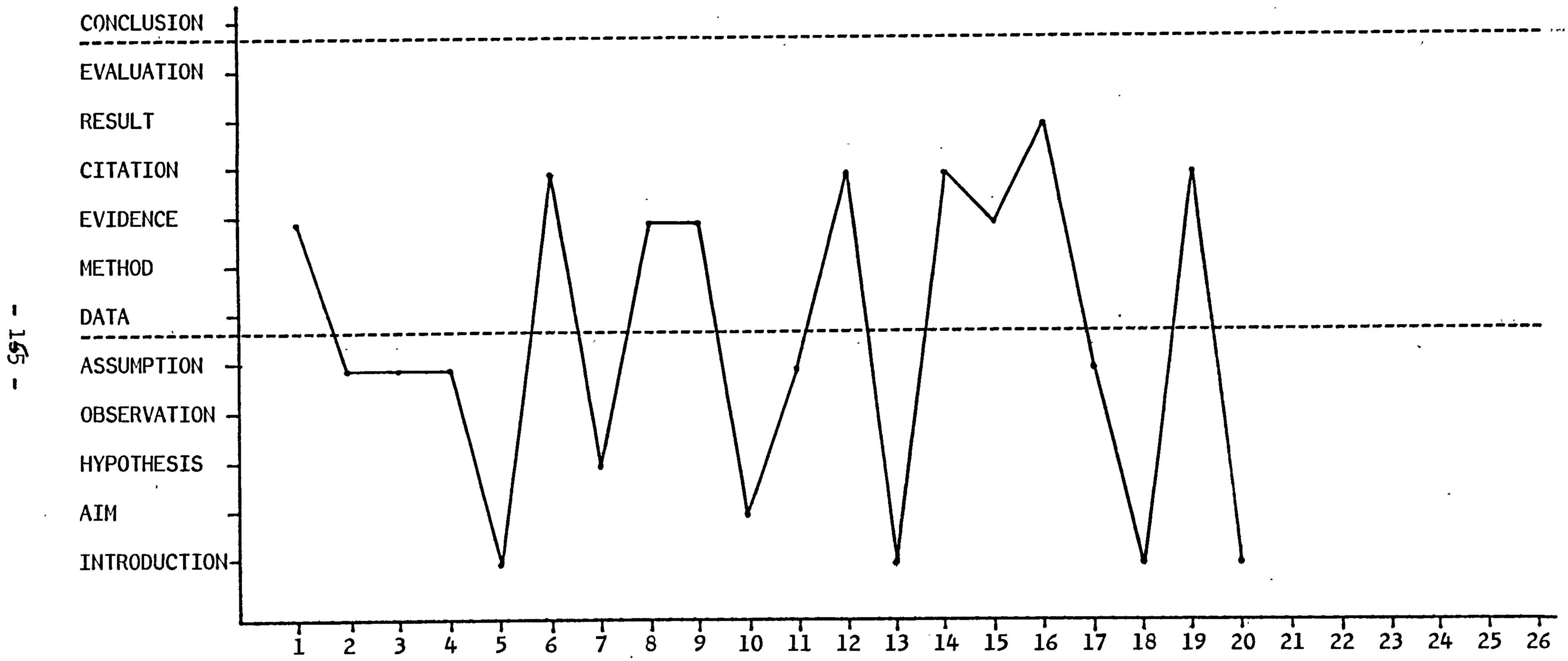
TEXT NO: 7


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

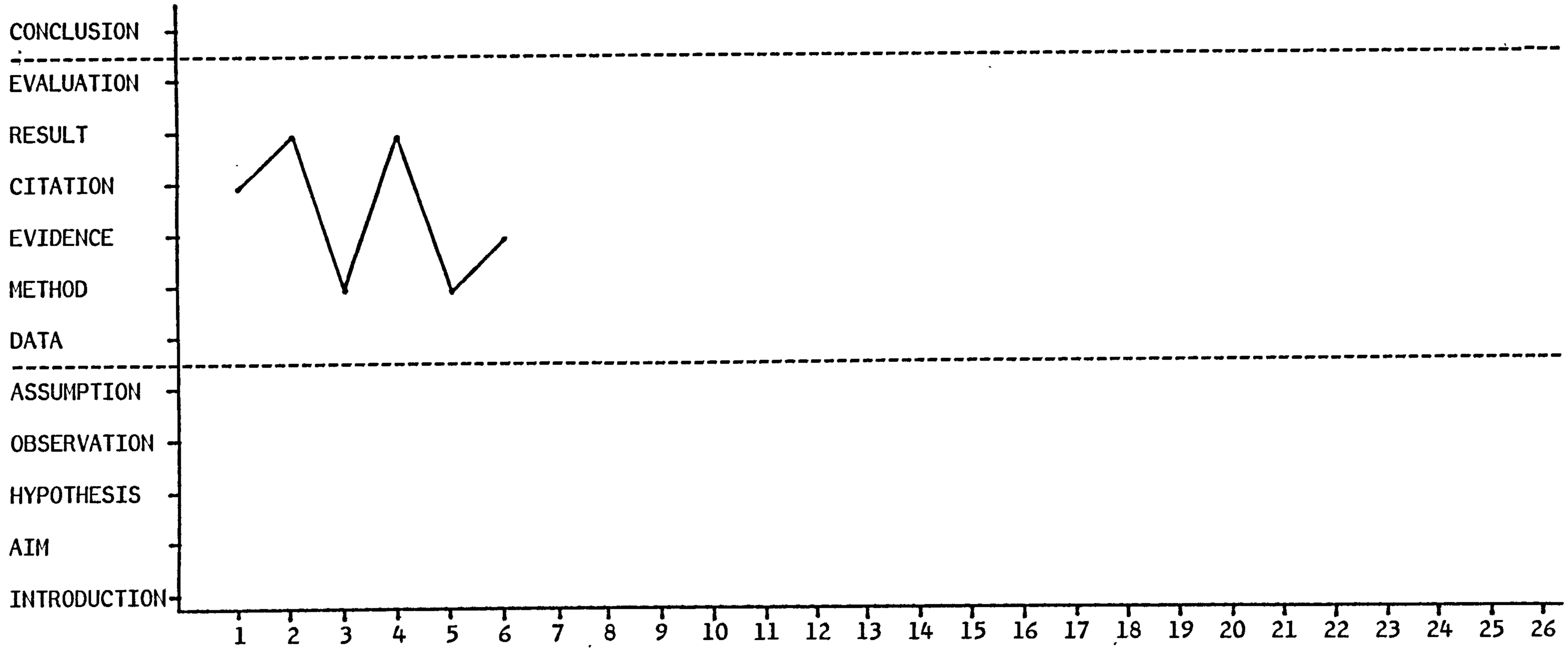
TEXT NO: 8


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

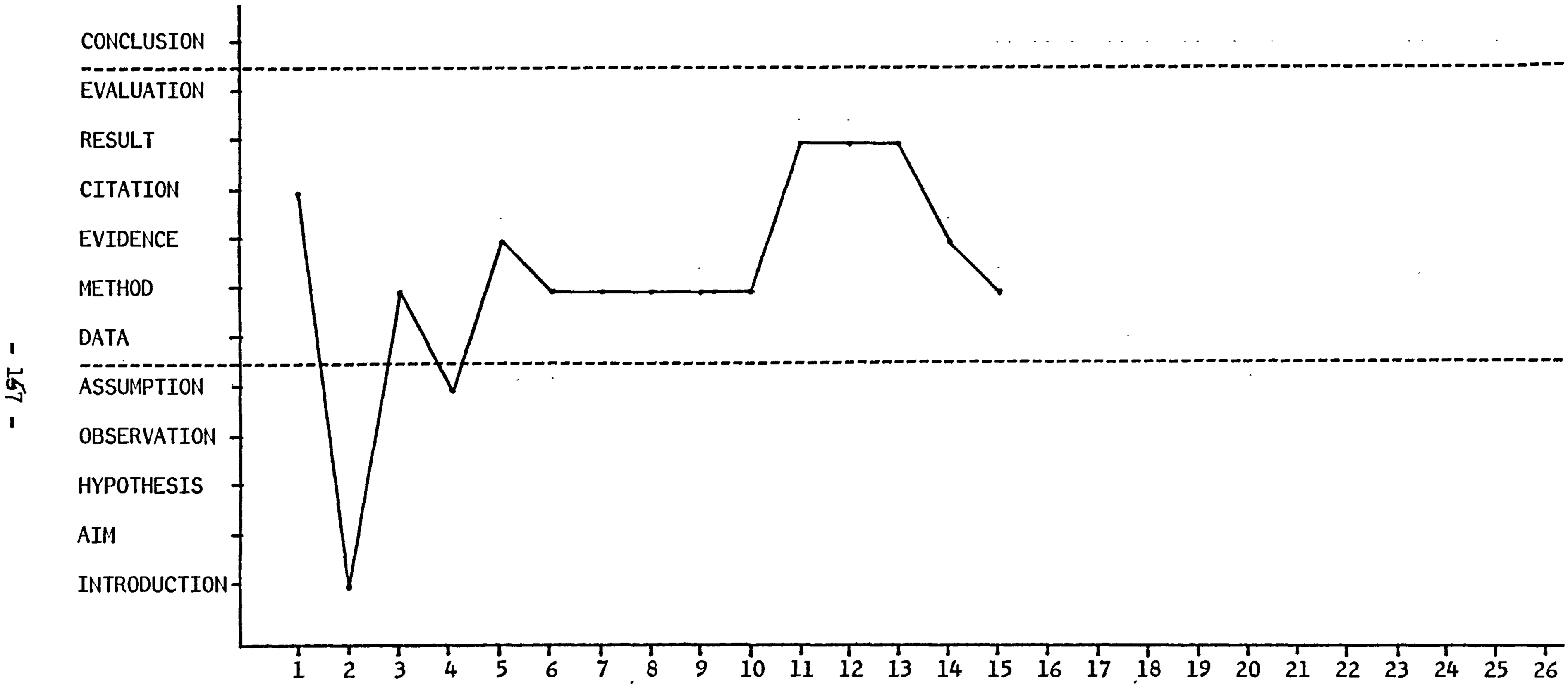
TEXT NO: 9


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

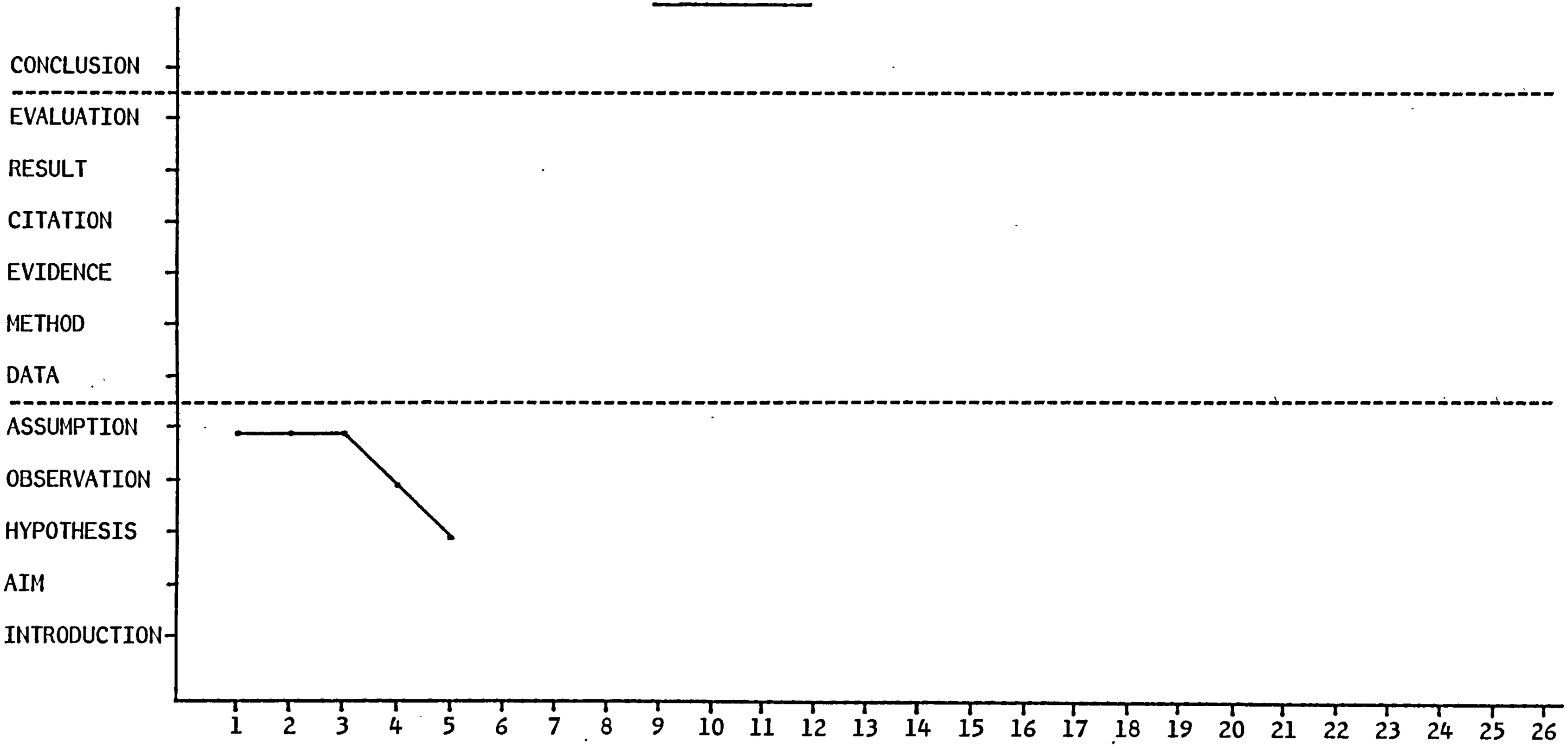
TEXT NO: 10


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

TEXT NO: 11


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

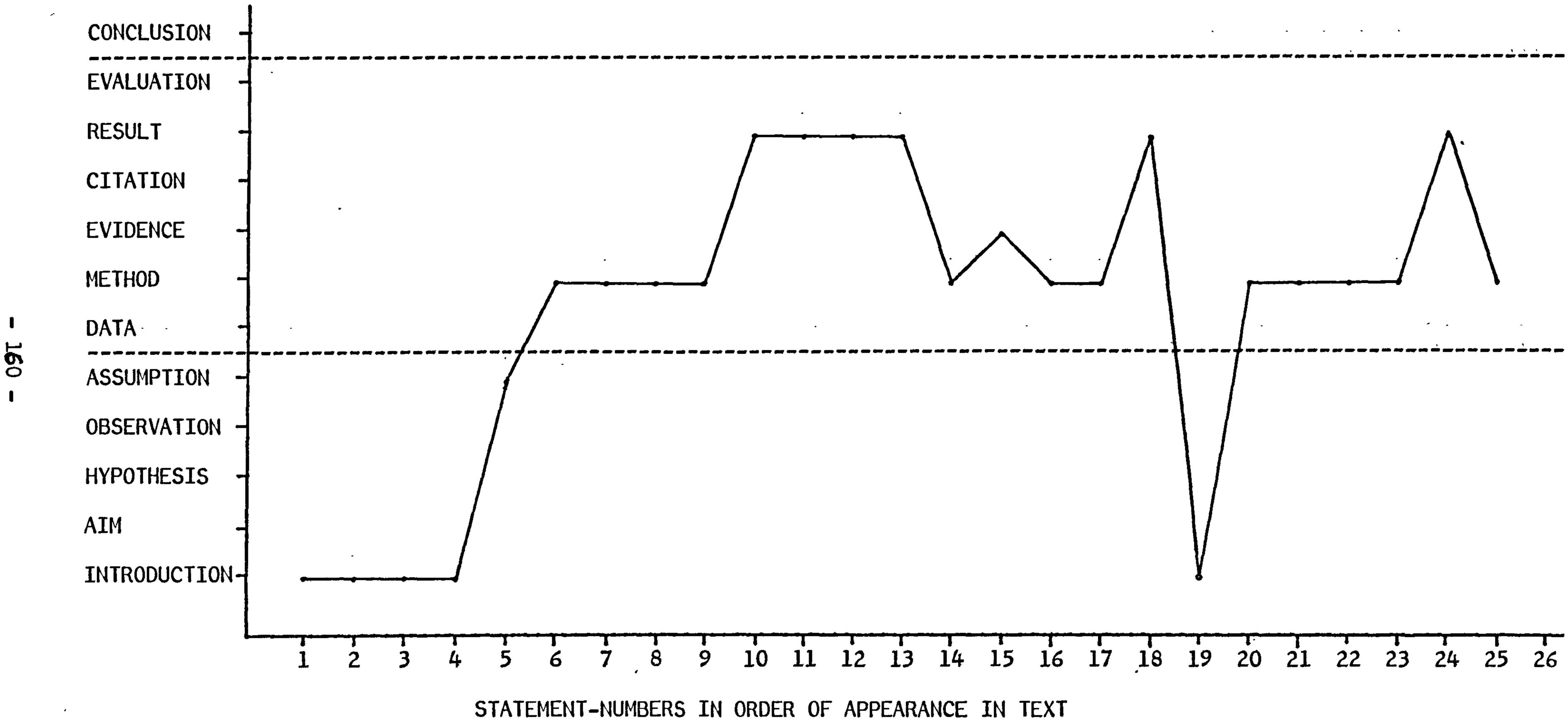
TEXT NO: 12


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

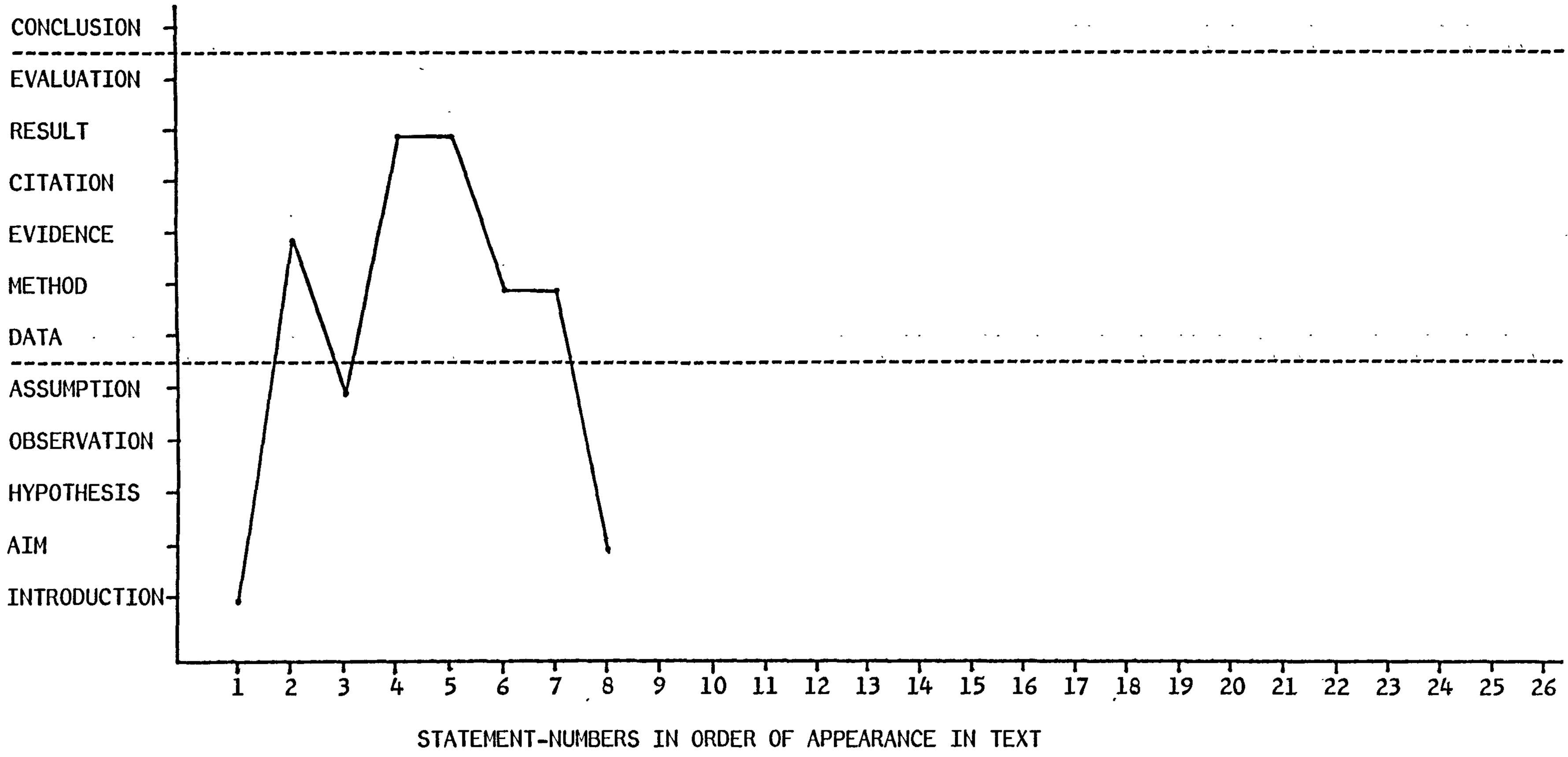
TEXT NO: 13


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

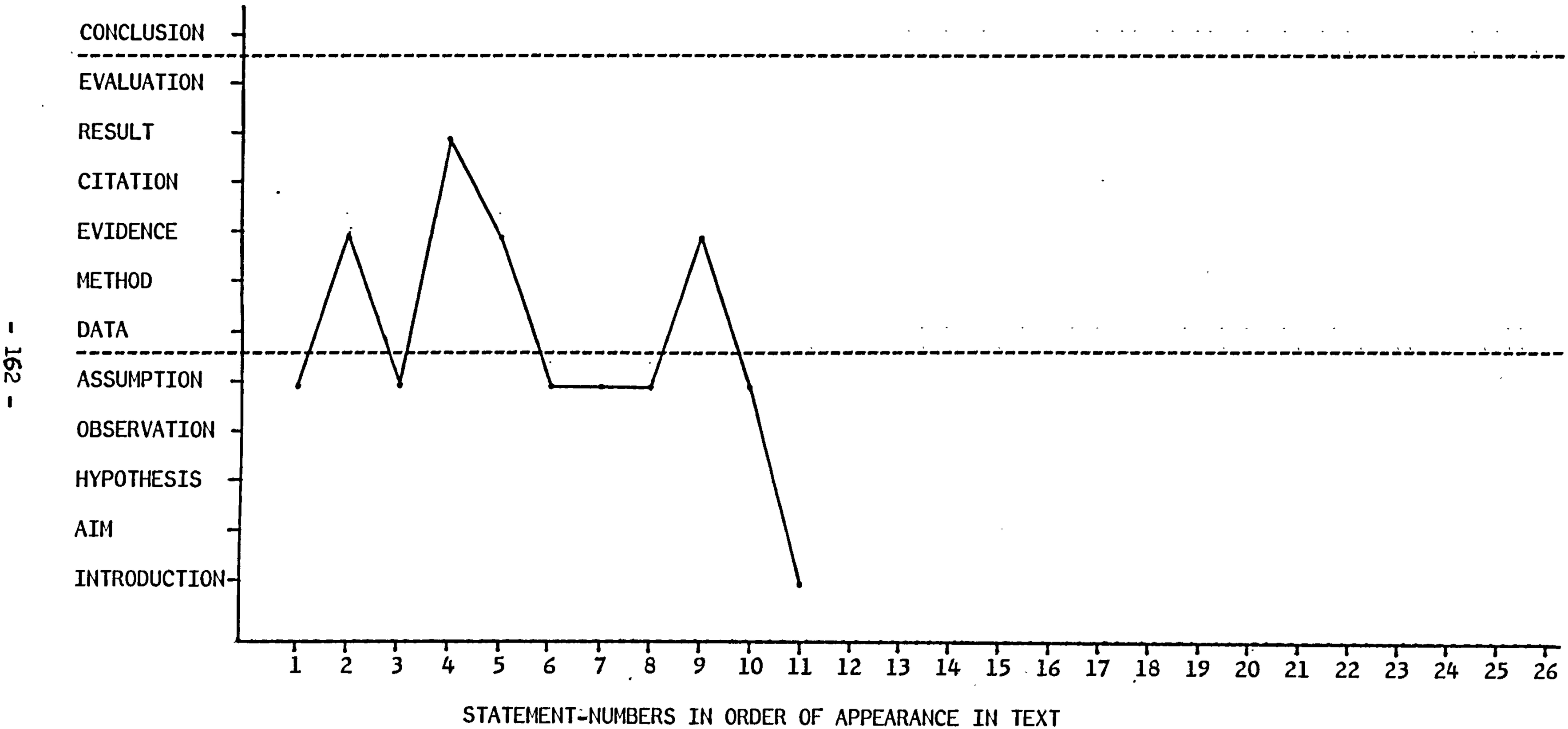
TEXT NO: 14


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

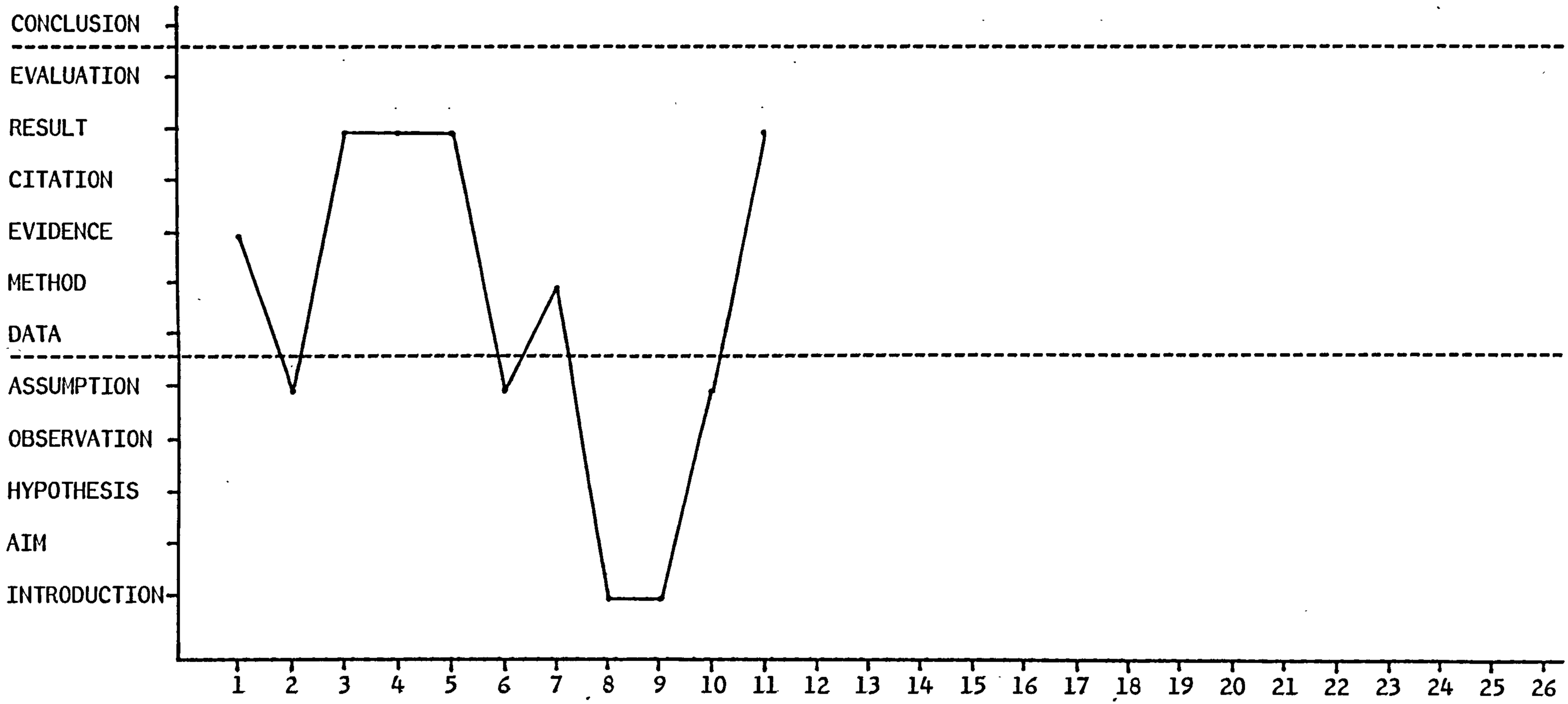
TEXT NO: 15


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

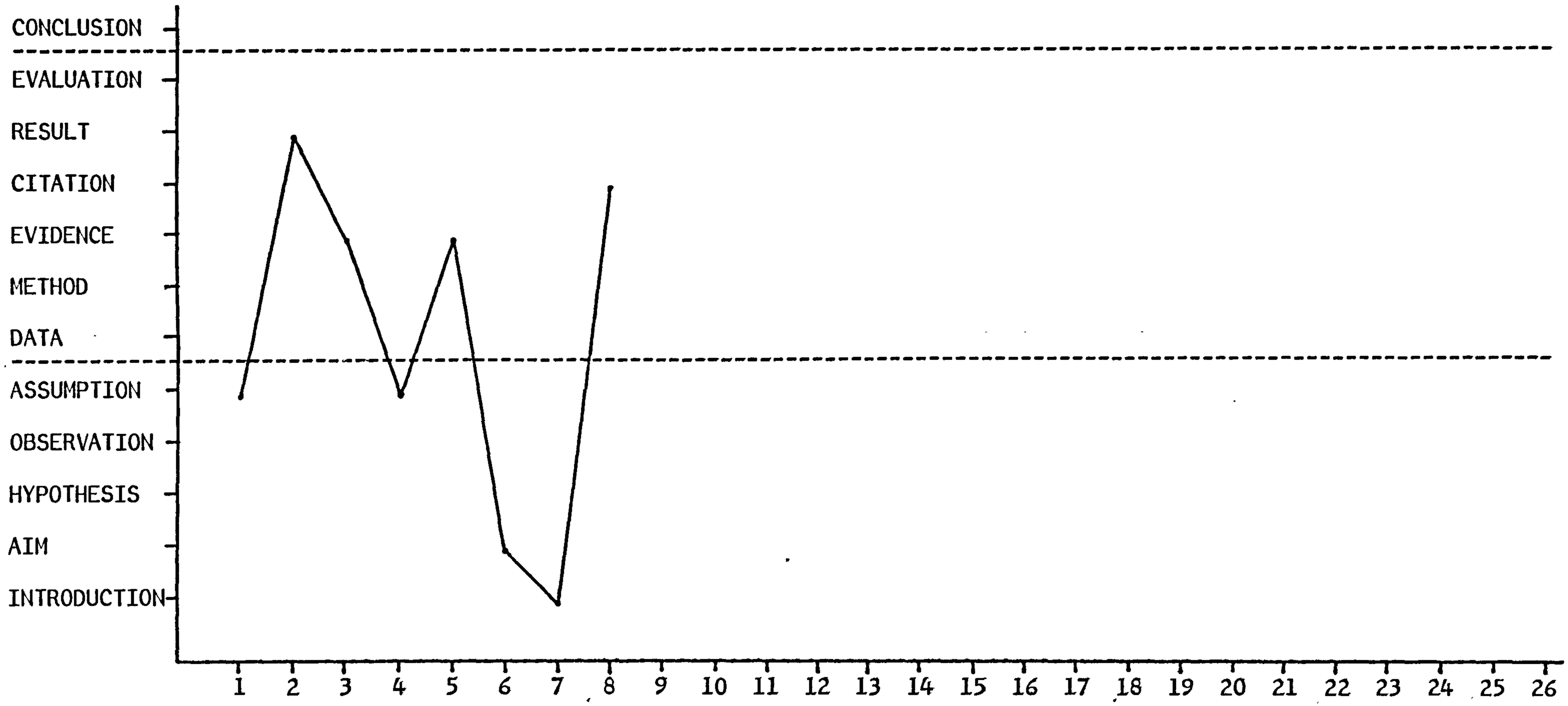
TEXT NO: 16


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

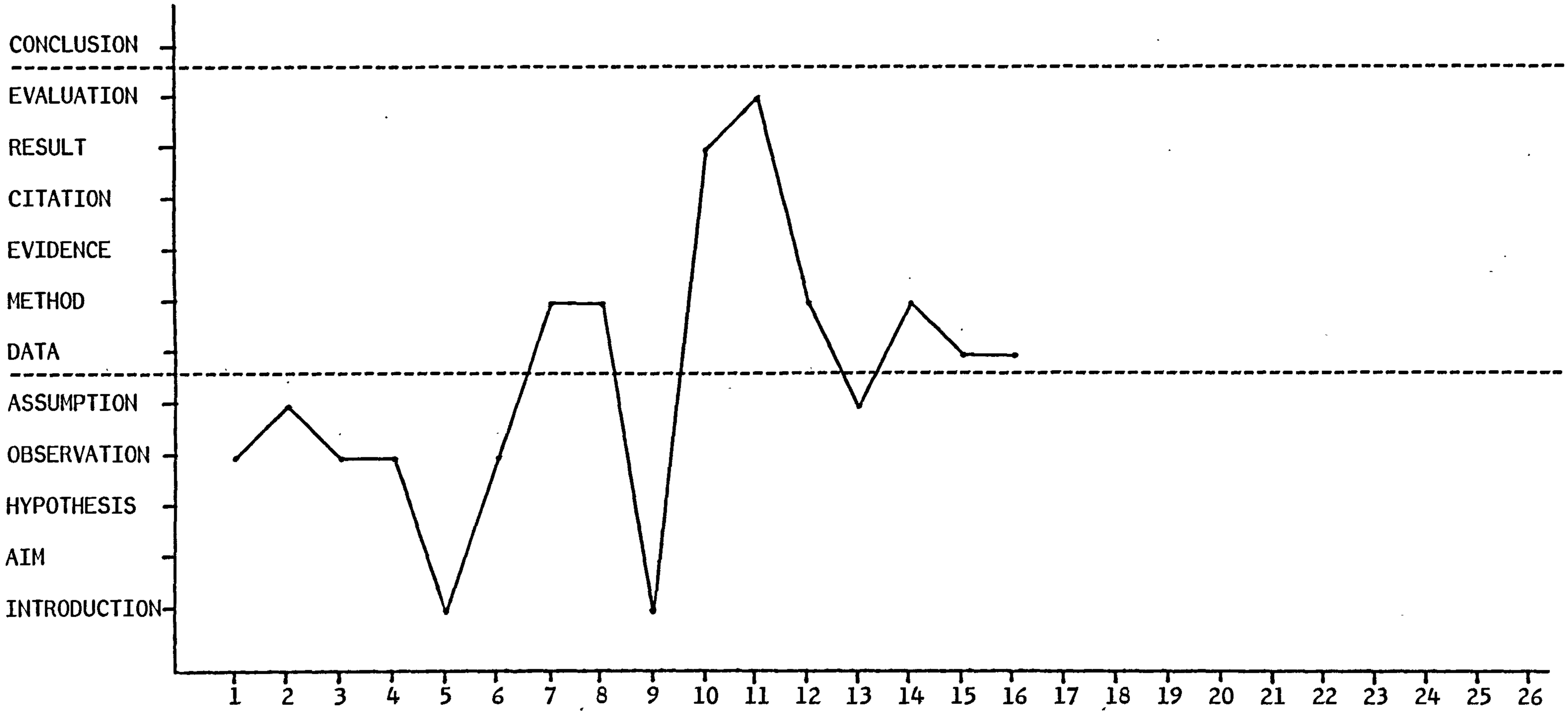
TEXT NO: 17


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

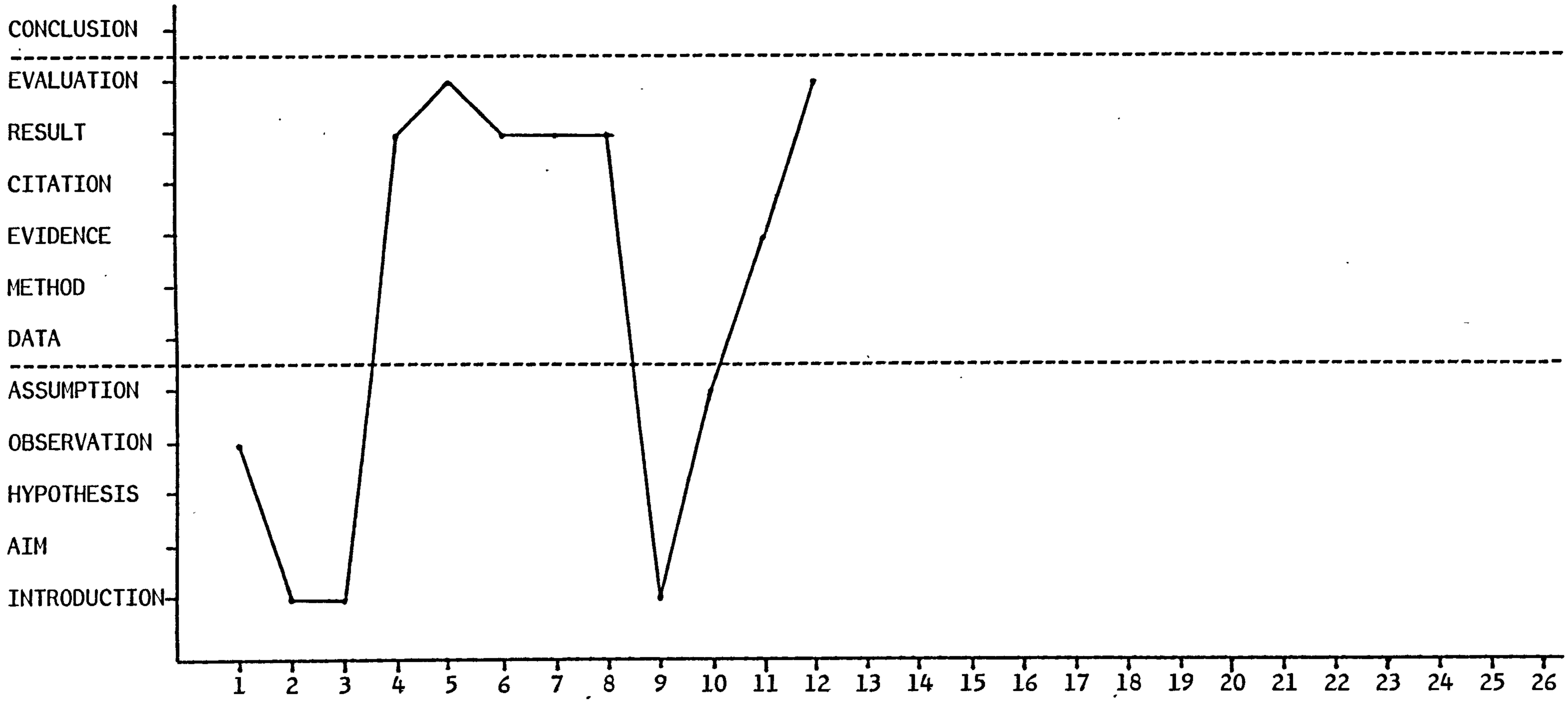
TEXT NO: 18


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

TEXT NO: 19


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

TEXT NO: 20


CONCLUSION
EVALUATION
RESULT
CITATION
EVIDENCE
METHOD
DATA
ASSUMPTION
OBSERVATION
HYPOTHESIS
AIM
INTRODUCTION
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
STATEMENT-NUMBERS IN ORDER OF APPEARANCE IN TEXT

APPENDIX C

SAMPLE TEXT FOR THE EXPERIMENTS IN PART II OF THE PILOT STUDY

Following is the text used by three subjects in the text analysis experiment mentioned in Part II of the Pilot Study. I am grateful to the author and editor of the Journal of Informatics for permission to reprint this text.

THE NEED FOR PSYCHOLINGUISTIC TESTING TO SOLVE PRACTICAL PROBLEMS IN INFORMATION RETRIEVAL

E Michael Keen
(Department of Information Systems Studies, College of Librarianship Wales)

Information retrieval systems usually provide, in response to a search, a visually scanned output in the form of index entries and/or document surrogates. Thus the searcher has to read this output and understand it, whether it is the entries in a printed subject index, the card notifications of a current awareness service, a display on a VDU screen, the printout of a computer terminal, or some other kind of output. The problem of intelligibility and comprehension appears to be at its worst when considering just the index entry portion of a printed index, without the help of a document surrogate such as a title or an abstract. This may be seen in Table 1, where six index entry types are given.

Type 4 in Table 1 is three entries from one of the subject index issues to *Psychological Abstracts*. Were this to be read as normal text no entry would read as a fully formed sentence, none would fit either active or passive order, and on average 25% of the possible function words are absent. Types 1, 3, and 5 differ in that they do have a complete set of function words, and type 1 reads in a sentence-like

way, here as active order phrases. Type 2 dispenses with all function words but retains sentence-like order, and type 6 (as 5) follows an order based on a scheme of contextual dependancy . All these represent well used index types, and EPSILON (1) has conducted experiments in which people search such indexes, and measures of quality and time are applied.

Quickly and accurately understanding the meaning of such index entries is a vital part of the search process, so that only those documents that look as though they will be relevant are followed up further. Thus it is a matter of correct relevance prediction. In linguistic terms the problem appears to be one of the comprehension of surface structures that are not usually very sentence-like in a situation akin to reading, but not one where speaking or listening are involved. How do people process such entries? Does type 4 provoke a transformation of *Psycholinguistics, survey of* ... into a possible sentence order, such as *Survey psycholinguistics* ... ? Do people using types 2 and 6 "invent" suitable function words? Some possible beginnings of answers to these questions have been uncovered in the EPSILON work (2), in which tape-recorded verbalised searches were made.

The view that I.R. must turn its attention to psycholinguistics was expressed recently (3). If this is correct the next step is to find out whether psycholinguistics researchers have tackled these problems. Is anyone currently doing work remotely related? Who has experience of setting up experiments involving people reading material, rather than making responses to audio cues? Have there been any psycholinguistic studies directed to rather practical text content problems at the surface structure level, rather than the fundamental linguistic problems of deep structure? These are the questions that now need to be posed to people in the field (4).

The immediate goal of interest in this area is the proper solution of completely practical problems in the design of indexes and information retrieval systems. There is no point at all in trying to discover a "transformational grammar" of index entries if linguists cannot yet do that for other situations of speech and reading. But a range of practical problems that are unresolved does exist, and such problems cannot ever be thoroughly understood unless some basic

questions about comprehension in the search process are tackled. For example, many indexes delete terms that are to be repeated in several entries, presumably to give convenient "packets" of entries in the display panorama (5), or to save space, paper, and printing cost. But how does this practice affect intelligibility and the correct prediction of relevance? Again, some indexes utilise a modifier line that has been carefully devised by an indexer to be both more readable and more accurate than a document title, resulting in an entry that is longer than the average title. But, does this approach decrease true intelligibility because it must cope also with remembering the search query, and perhaps also a "lead" term ten lines up the page that has not been repeated?

Table 2 sets out examples of four areas for linked theoretical and practical investigations, in each case the former leading to the latter. It may well be that direct work on these practical problems can be done given a suitable methodology, but when the time is ripe work must be done on the theory side so that proper explanations for practical result can be given, and universal principles established to apply to design problems.

The EPSILON work has been measuring the performance of controlled text indexes under laboratory conditions, including index entry types 1, 2 and 3 in Table 1, together with Articulated Term (no function words), Shunted Relational (as PRECIS, 27% function words), a Chain Procedure system, and a typographic variant of Rotated Term. The experience of attempting to devise valid test methods to accomplish this has lead to the linguistic considerations set out here. The text methodology still poses problems, and it would be unwise to claim that EPSILON has shown precisely how such tests can be carried out. There are at least four areas of competence required:

1) The implications of this being a problem of relevance predictability, and thus not being a case of mere comprehension of knowledge as presumably would be the case for educationalists investigating the teaching of reading. Relevance predictability in I.R. has not gone far yet (6).

2) The need for suitable laboratory and field test methods. The part-index scanning method being tried in EPSILON does have some considerable promise. Recording results unobtrusively at sufficient depth is a sub-problem.

3) The experimental design and statistics of such test methods warrant especial attention (7,8).

4) Any psycholinguistics research that has anything to offer to these problems must be absorbed and utilised.

Two questions remain. The first is whether these problems are of importance for I.R. systems that produce output (and permit searching) different from printed indexes. Table 3 gives some samples in which document titles are complemented by various kinds of descriptors and sentences in various layouts. The need to decide whether or not to obtain the full documents is pervasive to all I.R. systems that store surrogates only, so relevance prediction is vital in all such systems. The great variety in surrogate content, and indeed the variants shown in Table 3, suggest that producers are not at all clear about what is best, but even the most superficial comparisons are few (9).

The second question is whether this area should be tackled at all, and if it should, when. Obviously a field as small as information science cannot afford to support a lot of long term basic research, but it should tackle what is clearly its own province. A good look at psycholinguistics should show whether or not these problems are the province of I.R. Whether we do the research, or someone else does is not the most important matter: that our understanding of the theory should always be pushed to reasonable limits is essential so that improvements to the real world of information handling can be soundly introduced. The only reason for drawing back from this area would be if psycholinguists have tackled analogous problems, eg. with a straightforward text content and comprehension objective, and have failed to make any progress. So the next step is a cold hard look in that direction.

THE CLASSIFICATION OF STATEMENTS IN SCIENCE TEXT USING A GRAMMAR WHICH REFLECTS THE CONVENTIONAL FORM FOR REPRESENTING EMPIRICAL ARGUMENTS.

Instructions to participants:

This experiment is an opportunity for interested individuals to attempt the classification of statements in science text using the grammar provided below.

Use the following set of labels and rules to classify statements in the text on the next page. Any statement which you consider to be unclassifiable using this grammar should be underlined and an explanatory note made in the margin or on another piece of paper. Put a sequential number before each individual statement in the text. On the graph provided, mark the co-ordinate for each individual classified statement.

The tape recorder is provided for you to give a 'running commentary' of your analysis. Please try to talk continuously and discuss any problems you are having at the time.

Thank you for your interest.

The Grammar:

Labels: (these will be described to you verbally in more detail)

empirical argument	::=	Phase One Phase Two Phase Three
Phase One	::=	introduction aim hypothesis assumption observation
Phase Two	::=	method citation evidence result evaluation
Phase Three	::=	conclusion

Rules:

1. Only use one label for each statement. Labels may be used more than once.

2. As a general rule, treat sentences as statements.

3. Only classify assertive statements or propositions. NOT questions, for instance.

4. Identify statement-types by indicative words or phrases. That is, the phrase 'the aim of this paper is to ...', would indicate the use of the label aim . If no direct relationship between words in the statement with labels in the grammar exists, use thesaural-type relationships to determine the classification. For example, the term 'outset' is related to introduction .

5. Where anaphora or other reference from one statement to another occurs, try to classify the statement in question as a separate entity from the other statements around it by substituting say, pronouns with the noun being referred to.

APPENDIX D

SAMPLE TEXT FOR THE EXPERIMENTS IN TEXT ANALYSIS BY SCIENTISTS AND NON-SCIENTISTS

Following is the text used by the scientists and non-scientists in the text analysis experiments mentioned in sections 4.5 and 4.6. I am grateful to the author and editor of the Journal of the British Astronomical Association for permission to reprint this text.

JOURNAL OF THE BRITISH ASTRONOMICAL ASSOCIATION
Vol. 86 (1), Dec. 1975, pp. 44-46

THE MOON ILLUSION

Peter Cope

For countless years it has been noted that when the Moon is in the vicinity of the horizon it appears larger than when it has attained a greater altitude. In country areas where astronomical knowledge is, in general, deeply rooted amongst the countless 'old wives' tales, this is a familiar curiosity, but in our cities, a clear view of the horizon is becoming more and more difficult to obtain and consequently, many observers from such locales go through many years without ever observing this effect. What should be noted, however, is that the illusion is not solely confined to the Moon; proximity to the horizon also increases the size of the Sun and the area of constellations. This is an important point since it shows that every object in the heavens is similarly affected, but for our purposes here we shall deal only in the Moon for reasons of luminosity, size and visibility.

Despite the evidence of our eyes there is no real physical alteration in the size of the Moon, for we know it is impossible for such a body to swell and contract periodically merely by virtue of its position relative to an observer on Earth. We are, then, forced to conclude that it is a result of atmospheric distortions or something due to our own bodies.

That our atmosphere is in no way to blame can be demonstrated by a simple photographic test. A first photograph is taken when the Moon is low and displaying the illusion, and a second when the Moon is higher. The only stipulation is that the exposure is so arranged that both images on the resultant negatives have the Moon clearly defined and not extending into the sky. By very accurate measurement and comparison of the two images, no difference in size would be noted.

Our field is thus narrowed to a region of the head. The most natural assumption to make, in my case at least, because of the coupling of size change with altitude, is that the attitude of the eye is in some way to blame: i.e. changing the viewing angle causes eye distortion. Had I carried out research on the subject of the eye more thoroughly, I would have found that the whole of this experiment was so obviously incorrect that it should never have been carried out in the first place, but learning by my mistakes, I did start an elaborate sequence of experiments. Over a period of about three lunations, I spent my clear evenings either bent over double or lying flat on the muddy earth using various multiple mirror configurations which would reflect an image of the risen Moon against the horizon and vice versa. Needless to say, the result was total failure, and some advice from a knowledgeable biologist later crystallized my doubts about this particular theory. I wished I had discovered this earlier; my old monitoring planetarium raised a few eyebrows to put it mildly.

At the same planetarium a new line of thought developed. Most planetaria have specially cut horizons about the circumference. It seemed to me that considerable importance was attached to the horizon, and the real horizon could, then, be crucial to the formation of the illusion.

Acting on a hunch, not really knowing what conclusions could be drawn, I began estimating the size of the setting Moon when against different horizons. My methods of gauging the size of the Moon were admittedly crude and unscientific (the size being described as 'big', 'bigger' or 'biggest') but they were certainly effective. I am fortunate in having within a couple of hundred metres of my home a varied horizon ranging from a few metres above rooftops to many kilometres across the plains of Windsor and Slough.

The results are shown below:

1. Over rooftops: normal size;
2. Over more distant rooftops (200 m±) 'big';
3. Over distant hillock (1.5 km) 'bigger';
4. Over distant horizon (10 km+) 'biggest'.

It appears from these results that the more distant the horizon is, the larger the Moon appears. I did not, though, trust the evidence of my own eyes and asked several friends, with no technical knowledge of the Moon illusion, to duplicate my tests. Out of the 40 observations of the 10 'guinea pigs', the results kept fairly well within the scale above.

What I am now writing is a simplification of my original explanation, which, as it had just been prepared, was lacking details and rather sketchy. The final explanation involves a subtle combination of two effects, which I shall briefly outline.

Objects on the Earth are, for the most part, never seen in such a position as to draw comparisons with the Moon. The time when this does happen is when the Moon is setting and rising. At these times a link is formed between these two objects. The results of the above experiment showed how distance affected size. Some imagination is required at this point: imagine a close-by tree with the Moon above it. The former would dwarf the latter, and the brain interprets the Moon as a small although distant object. Secondly, imagine some trees upon a distant horizon; now the situation is reversed, and it is the distant Moon that is larger than the similarly (!) distant treescape. Our habits of thought have constantly dealt with houses

and trees and we are familiar with their size. The Moon is then compared with these objects of known dimensions and appears larger above the familiar objects.

The second effect concerns our stereoscopic vision, and compared with the first, is negligible in extent. When we have nearby trees, taking the case above, we can tell unmistakably which is the nearer, but when it comes to a distant horizon (many miles away) the decision is much more difficult to make. The Moon seems to have risen directly from the skyline. Unfortunately, there is no way of recreating a distant horizon for recognition from a house roof. By moving my vantage point I could produce moon-sets, but on not one occasion had I the slightest illusory effect. It seems that there is no substitute for the real thing.

At its most basic state the Moon illusion itself can be attributed to the linkage of Moon with horizon, which does not occur except when the Moon is rising or sinking in the sky. The change in size apparently displayed increased with a proportional increase in horizontal distance.

It is from the aforementioned subtle combination of effects that the Moon illusion and horizon/size relationships probably stem; undoubtedly others could explain it far more concisely and less heavy-handedly, but I hope that I have provided sufficient interest to merit some further investigations.

CLASSIFYING STATEMENT-TYPES IN TEXT USING A SET OF SEMANTIC DESCRIPTORS

Instructions to participants:

Please read the attached text and try to classify each individual statement using one of the semantic descriptors listed below. In most cases a statement will be a complete sentence. If any statement is in your opinion unclassifiable using the descriptors below, enter a '?' on the data sheet provided. If any statement is ambiguous, enter a '*' followed by the descriptors which could be used.

Category	Semantic Descriptor	Code
Category One	Introductory	1A
	Aim	1B
	Hypothesis	1C
	Assumption	1D
	Observation	1E
Category Two	Method	2A
	Evidence	2B
	Citation	2C
	Result	2D
	Evaluation	2E
Category Three	Conclusion	3A

Please put a sequential number in front of every statement located and classified or not.

APPENDIX E

SUMMARIES OF THE 'MOON ILLUSION' TEXT

The summaries given below were produced by two groups of subjects in the experiment described in section 4.7 of Chapter Four. Numbers preceeding statements denote classifications for each from the grammar. Classification of statements from these sequential numbers appears after each summary.

EXPERIMENTAL GROUP

Subject One:

(1) The paper is concerned with the illusion that the Moon, (and also other bodies but the discussion is confined to the Moon because of ease of observation) appears larger the lower in the sky it is.

(2) We assume the Moon's size and distance is effectively constant. (3) Atmospheric distortion can be eliminated by careful comparison of photos of the Moon at different heights in the sky. (4) It is known that apparent size does not change with observer's viewing angle. (5) The illusion could therefore be due to the nature of the horizon; (6) this hypothesis was tested using several observers, all of whom obtained similar results: (7) the further the horizon, the larger the Moon appeared.

(8) The explanation is that at rising and setting the Moon and horizon are viewed in conjunction and the apparent size of the moon judged against the known, because familiar, size and distance of objects on the horizon. (9) This is compounded by the effects of stereoscopic vision which facilitates judgement of the distance of near rather than far objects.

(1) Introduction
(2) Assumption
(3) Method
(4) Evidence
(5) Evaluation
(6) Method
(7) Result
(8) Conclusion
(9) Conclusion.

Subject Two:

(1) This paper investigates the illusion of increased size of the Moon when in the vicinity of the horizon. (2) The reason for the illusion is thought not to be the attitude of the eye. (3) Observations of apparent Moon size indicate that Moon size appears to be increased as seen against increasing distant skylines. (4) The author interpretes this as the readjustment by the brain of Moon size with the size of familiar objects. (5) A more slight effect is that of stereoscopic vision and the difficulty of judging relative distances at the horizon.

(1) Introduction
(2) Assumption
(3) Observation
(4) Evaluation
(5) Result

Subject Three:

(1) When the Moon is in the vicinity of the horizon it appears larger than when it has attained a greater altitude, this illusion is not confined to the Moon. (2) Reason tells us that there is no real physical alteration in the size of the Moon. (3) By the means of a photographic test, (4) altitude was found not to affect Moon size. (5) Likewise tests on viewing angle showed this did not cause this distortion either. (6) Considerable importance is attached to the horizon in most planetaria - so perhaps this could be crucial to the formation of the

illusion. (7) By examining the Moon over differing horizons it was found the more distant the horizon is, the larger the Moon appears. (8) Objects on the Earth, are generally never seen in such a position as to draw comparisons with the Moon. (9) Distance affects size, we are used to dealing with houses and trees but unused to comparing the Moon with those objects of known dimensions. (10) No illusory effect could be produced artificially by the use of Moon sets.

(11) The apparent change in size of the Moon when rising or sinking increases with a proportional increase in horizontal distance.

(1) Observation
(2) Assumption
(3) Method
(4) Result
(5) Result
(6) Data
(7) Method
(8) Evidence
(9) Data
(10) Result
(11) Conclusion

Subject Four:

(1) Moon in the vicinity of horizon appears larger.
(2) There is no physical alteration in the size of the Moon.
(3) Assumption: atmosphere[1] on our body to blame.[2]
(4) Elimination of assumption 1. (test)
(5) Therefore: assumption 2 must be true.
(6) Assumption 2 eliminated (experiments proved futile).
(7) New hypothesis: horizon is responsible for the phenomenum.
(8) Experiments confirm the hypothesis: the more distant the horizon, the larger the Moon.
(9) Explanation (scientific): stroboscopic vision; comparison of familiar with unfamiliar.
(10) Conclusion: moon illusion is attributed to the linkage of Moon with horizon.

(1) Hypothesis
(2) Assumption
(3) Assumption
(4) Result
(5) Evaluation
(6) Result
(7) Hypothesis
(8) Data
(9) Method
(10) Conclusion

Subject Five:

(1) The moon, and all other celestial objects, appear larger when on the horizon. (2) Established photographically not to be an atmospheric effect. (3) Therefore must be psychological. (4) Experiments were then carried out estimating the size of the moon against horizons at different distances. (5) It was found that the further the horizon, the larger the moon appeared to be. (6) Hence it was concluded that the illusion is a problem of the comparison of the size of the moon with distant objects, and (7) thus it is confined to times when the moon is low enough for these comparisons to be made, i.e. when the moon is rising or setting.

(1) Observation
(2) Result
(3) Conclusion
(4) Method
(5) Result
(6) Conclusion
(7) Conclusion

CONTROL GROUP

Subject One:

(1) It has been observed that the Moon appears larger when on the horizon. (2) By experimentation it was established that the Moon seemed to be larger the further away from the horizon one observed it. (3) The conclusion from this evidence was that the illusion of the Moon's size was caused by a difference in size of other objects on the horizon relative to the distance from the horizon that one observed the phenomenum.

(1) Observation
(2) Method
(3) Conclusion

Subject Two:

(1) The Moon appears larger on the horizon than when high in the sky. (2) Photographs showed that this was not due to any atmospheric conditions and therefore must be a psychological or optical illusion. (3) Experiments which recorded size differences at greater and less distances from the horizon showed that the Moon appeared larger at greater distances from the horizon. (4) This was due to the diminishing size of other objects on the horizon which the Moon's size could be compared with nearer the horizon.

(1) Observation
(2) Result
(3) Method
(4) Conclusion

Subject Three:

(1) Celestial objects, including the Moon, are often observed to be larger on the horizon. (2) If the size of the Moon is compared with other earthly objects near the horizon, then the comparison is repeated at increasingly greater distances, the Moon appears to get even bigger. (3) This was apparently, because the earthly objects themselves became smaller. (4) Such comparisons can only occur when the Moon is low on the horizon.

(1) Observation
(2) Result
(3) Evaluation
(4) Conclusion

Subject Four:

(1) The Moon, compared with other objects on the horizon appears larger at a distance than close to the horizon. (2) Experiments to test this observation were made and showed that this was so. (3) The reason for this illusion was that the further away one got from the horizon, the smaller the objects became,. thus appearing to increase the size of the Moon.

(1) Introduction
(2) Method
(3) Result

Subject Five:

(1) It was observed that the Moon appeared larger on the horizon than in the sky. (2) The method of testing this phenomenum was to take photographs of it first to show that it was not an atmospheric effect. (3) The next method was to observe the Moon at variable distances from the horizon and compare its size at those points. (4) The result was that it looked larger at a distance. (5) The conclusion was that the Moon appears larger at a distance because of the relatively decreasing size of other objects on the horizon.

(1) Observation

(2) Method

(3) Method

(4) Result

(5) Conclusion

APPENDIX F

CODE LISTINGS FROM COMPUTER PROGRAMS

Following are the code listings for the two computer programs discussed in Chapter 5.

SUMMARY GENERATION PROGRAM

```
LIST
  10  REM PROGRAM-ID = ASR(ABSTRACT GENERATOR)  -  VERSION #1
  20  REM AUTHOR PHILIP J. SALLIS,  27/9/78
  30  LET A=0
  40  LET A$="PRINCIPAL HYPOTHESIS"
  50  LET C$="PRIMARY AIM"
  60  LET E$="INTRODUCTORY ASSUMPTION"
  70  LET G$="FACT/DATA"
  80  LET I$="CITATION"
  90  LET K$="METHOD"
 100  LET M$="RESULTS"
 110  LET O$="METHODOLOGICAL ASSUMPTION"
 120  LET Q$="PRIMARY CONCLUSION"
 130  LET S$="DEDUCTIVE CONCLUSION"
 140  LET U$="INDUCTIVE CONCLUSION"
 150  REM
 160  PRINT
 170  PRINT "THIS PROGRAM GENERATES ABSTRACTS FROM DATA SUPPLIED"
 180  PRINT "BY THE AUTHORS OF DOCUMENTS.  QUESTIONS ARE ASKED BY"
 190  PRINT "THE PROGRAM, WHICH THEN GENERATES AN ABSTRACT BASED"
 200  PRINT "ON THAT DATA AND A GRAMMAR WHICH IS HELD IN THE SYSTEM."
 210  REM
 220  REM LOOP FOR QUESTIONING THE USER
 230  LET A=A+1
 240  IF A>11 THEN 360
 250  IF A=1 THEN 410
 260  IF A=2 THEN 440
 270  IF A=3 THEN 470
 280  IF A=4 THEN 500
 290  IF A=5 THEN 530
 300  IF A=6 THEN 560
 310  IF A=7 THEN 590
 320  IF A=8 THEN 620
 330  IF A=9 THEN 650
 340  IF A=10 THEN 680
 350  IF A=11 THEN 710
 360  REM
 370  REM BRANCH TO PRINT-OUT THE ABSTRACT
 380  GOTO 1330
 390  REM
 400  REM TEST FOR CONTENTS OF Z$
 410  LET Z$=A$
 420  GOSUB 770
 430  GOTO 230
 440  LET Z$=C$
 450  GOSUB 770
 460  GOTO 230
 470  LET Z$=E$
 480  GOSUB 770
 490  GOTO 230
 500  LET Z$=G$
 510  GOSUB 770
 520  GOTO 230
 530  LET Z$=I$
 540  GOSUB 770
 550  GOTO 230
 560  LET Z$=K$
 570  GOSUB 770
 580  GOTO 230
 590  LET Z$=M$
 600  GOSUB 770
 610  GOTO 230
 620  LET Z$=O$
 630  GOSUB 770
 640  GOTO 230
 650  LET Z$=Q$
 660  GOSUB 770
 670  GOTO 230
 680  LET Z$=S$
 690  GOSUB 770
 700  GOTO 230
 710  LET Z$=U$
 720  GOSUB 770
 730  GOTO 230
 740  REM
 750  REM
 760
 770  REM GRAMMAR QUESTION SUBROUTINE
 780  PRINT
 790  PRINT "DOES YOUR ARTICLE HAVE "Z$
 800  PRINT "ANSWER 'Y' OR 'N':"
 810  PRINT
 820  INPUT Y$
 830  IF Y$="N" THEN 890
 840  PRINT "ENTER THE "Z$" IN NO MORE THAN ONE LINE:"
```

```
 850  PRINT
 860  INPUT X$
 870  GOSUB 940
 880  GOTO 910
 890  LET X$=" "
 900  GOSUB 940
 910  RETURN
 920  REM
 930  REM
 940  REM STORE TEXT SUBROUTINE
 950  REM
 960  IF A=1 THEN 1080
 970  IF A=2 THEN 1100
 980  IF A=3 THEN 1120
 990  IF A=4 THEN 1140
1000  IF A=5 THEN 1160
1010  IF A=6 THEN 1180
1020  IF A=7 THEN 1200
1030  IF A=8 THEN 1220
1040  IF A=9 THEN 1240
1050  IF A=10 THEN 1260
1060  IF A=11 THEN 1280
1070  REM
1080  LET B$=X$
1090  GOTO 1300
1100  LET D$=X$
1110  GOTO 1300
1120  LET F$=X$
1130  GOTO 1300
1140  LET H$=X$
1150  GOTO 1300
1160  LET J$=X$
1170  GOTO 1300
1180  LET L$=X$
1190  GOTO 1300
1200  LET N$=X$
1210  GOTO 1300
1220  LET P$=X$
1230  GOTO 1300
1240  LET R$=X$
1250  GOTO 1300
1260  LET T$=X$
1270  GOTO 1300
1280  LET V$=X$
1290  REM
1300  RETURN
1310  REM
1320  REM
1330  REM PRINT-OUT SUBROUTINE
1340  REM
1350  PRINT
1360  PRINT
1370  PRINT "##### ABSTRACT FOLLOWS:"
1380  PRINT
1390  PRINT
1400  IF B$<>" " THEN 1700
1410  LET B$=" "
1420  IF D$<>" " THEN 1720
1430  LET D$=" "
1440  IF F$<>" " THEN 1740
1450  LET F$=" "
1460  IF H$<>" " THEN 1760
1470  LET H$=" "
1480  IF J$<>" " THEN 1780
1490  LET J$=" "
1500  IF L$<>" " THEN 1800
1510  LET L$=" "
1520  IF N$<>" " THEN 1820
1530  LET N$=" "
1540  IF P$<>" " THEN 1840
1550  LET P$=" "
1560  IF R$<>" " THEN 1860
1570  LET R$=" "
1580  IF T$<>" " THEN 1880
1590  LET T$=" "
1600  IF V$<>" " THEN 1900
1610  LET V$=" "
1620  GOTO 1910
1630  REM
1640  REM BRANCHED-TO PRINT ROUTINES
1650
1660
1670
1680
```

```
1690  PRINT
1700  PRINT A$;" ";B$;"."
1710  GOTO 1420
1720  PRINT C$;" ";D$;"."
1730  GOTO 1440
1740  PRINT E$;" ";F$;"."
1750  GOTO 1460
1760  PRINT G$;" ";H$;"."
1770  GOTO 1480
1780  PRINT I$;" ";J$;"."
1790  GOTO 1500
1800  PRINT K$;" ";L$;"."
1810  GOTO 1520
1820  PRINT M$;" ";N$;"."
1830  GOTO 1540
1840  PRINT O$;" ";P$;"."
1850  GOTO 1560
1860  PRINT Q$;" ";R$;"."
1870  GOTO 1580
1880  PRINT S$;" ";T$;"."
1890  GOTO 1600
1900  PRINT U$;" ";V$;"."
1910  PRINT
1920  PRINT
1930  REM
1940  REM
1950  REM END OF PROGRAM
1960  END
```

RETRIEVAL PROGRAM

```
UNIVERSITY OF SOUTHAMPTON BASIC MARK 91  ON 27/09/78 AT 12.40.12

*OLD TXAN
OK
*LIST
   10  REM PROGRAM-ID = TXAN (TEXT ANALYSER).  AUTHOR = PHILIP J. SALLIS
   20  REM "INFORMATION RETRIEVAL SYSTEM SIMULATION" - VERSION 2 - 26/9/
78
   30  LET W=X=Z=0
   40  LET P$="COMPILER"
   50  LET Q$="MARBLES"
   60  LET R$="PIGS"
   70  LET S$="VEHICLES"
   80  PRINT
   90  PRINT "WELCOME TO THE IRS SIMULATION - PLEASE ENTER THE"
  100  PRINT "TERM OF YOUR CHOICE FROM THE FOLLOWING LIST:"
  110  PRINT
  120  PRINT P$,Q$,R$,S$
  130  PRINT
  140  INPUT A$
  150  IF A$=P$ THEN 220
  160  IF A$=Q$ THEN 250
  170  IF A$=R$ THEN 280
  180  IF A$=S$ THEN 310
  190  PRINT
  200  PRINT "*** ERROR IN INPUT - PLEASE RETYPE A TERM FROM THE LIST **
*"
  210  GOTO 130
  220  PRINT
  230  PRINT "REF: 1/1 - 'COMPILER FAULT TESTING'"
  240  GOTO 340
  250  PRINT
  260  PRINT "REF: 1/2 - 'GAMES WITH MARBLES'"
  270  GOTO 340
  280  PRINT
  290  PRINT "REF: 1/3 - 'THE FEEDING OF PIGS AND OTHER SUCH FUN'"
  300  GOTO 340
  310  PRINT
  320  PRINT "REF: 1/4 - 'HOW TO REPAIR MOTOR VEHICLES'"
  330  'REM CHOICE OF TITLE
  340  PRINT
  350  PRINT "DO YOU WANT TO INSPECT THIS DOCUMENT?  TYPE 'Y' OR 'N':"
  360  PRINT
  370  INPUT B$
  380  IF B$="Y" THEN 480
  390  PRINT "YOU HAVE TYPED 'NO', SO YOU CAN EITHER HAVE A PRINT"
  400  PRINT "OF THIS DOCUMENT, ANOTHER CHOICE FROM THE LIST GIVEN"
  410  PRINT "ABOVE, OR END THE PROGRAM.  TYPE '1', '2', OR '3'"
  420  PRINT "FOR WHICHEVER OF THOSE OPTIONS YOU WANT."
  430  PRINT
  440  INPUT Z
  450  IF Z=1 THEN 1330
  460  IF Z=2 THEN 90
  470  IF Z=3 THEN 1690
  480  PRINT
  490  PRINT "YOU HAVE INDICATED THAT YOU WOULD LIKE TO INSPECT"
  500  PRINT "THE DOCUMENT CONTAINING THE KEYWORD ";A$
  510  PRINT "MORE CLOSELY.  HERE IS A GRAMMAR THAT YOU CAN USE TO"
  520  PRINT "ASCERTAIN THE AUTHOR'S ARGUMENT IN THE DOCUMENT."
  530  PRINT
  540  PRINT "1    = HYPOTHESIS"
  550  PRINT "1A   = PRIMARY AIM"
  560  PRINT "1B   = INTRODUCTORY ASSUMPTION"
  570  PRINT "2    = FACT/DATA"
  580  PRINT "2A   = CITATION"
  590  PRINT "2B   = METHOD"
  600  PRINT "2C   = RESULTS"
  610  PRINT "2D   = METHODOLOGICAL ASSUMPTION"
  620  PRINT "3    = CONCLUSION"
  630  PRINT "3A   = DEDUCTIVE CONCLUSION"
  640  PRINT "3B   = INDUCTIVE CONCLUSION"
```

```
650  PRINT
660  PRINT "ENTER THE CODE NUMBER WHICH CORRESPONDS TO THE TYPE"
670  PRINT "OF STATEMENT YOU WISH TO DISPLAY FROM THE DOCUMENT"
680  PRINT
690  INPUT C$
700  PRINT
710  IF C$="1" THEN 830
720  IF C$="1A" THEN 860
730  IF C$="1B" THEN 890
740  IF C$="2" THEN 920
750  IF C$="2A" THEN 950
760  IF C$="2B" THEN 980
770  IF C$="2C" THEN 1010
780  IF C$="2D" THEN 1040
790  IF C$="3" THEN 1070
800  IF C$="3A" THEN 1100
810  IF C$="3B" THEN 1130
820  PRINT
830  PRINT "THE PRINCIPAL HYPOTHESIS OF THIS TEXT IS:"
840  GOSUB 1280
850  GOTO 1160
860  PRINT "THE PRIMARY AIM OF THIS TEXT IS:"
870  GOSUB 1280
880  GOTO 1160
890  PRINT "THE INTRODUCTORY ASSUMPTION OF THIS TEXT IS:"
900  GOSUB 1280
910  GOTO 1160
920  PRINT "THE FACT/DATA PHASE OF THIS TEXT BEGINS WITH:"
930  GOSUB 1280
940  GOTO 1160
950  PRINT "THE CITATIONS FROM THIS TEXT ARE:"
960  GOSUB 1280
970  GOTO 1160
980  PRINT "THE METHOD OF THIS ARGUMENT IS:"
990  GOSUB 1280
1000 GOTO 1160
1010 PRINT "THE RESULTS FROM THE WORK DESCRIBED IN THE TEXT ARE:"
1020 GOSUB 1280
1030 GOTO 1160
1040 PRINT "THE METHODOLOGICAL ASSUMPTIONS ARE!"
1050 GOSUB 1280
1060 GOTO 1160
1070 PRINT "THE PRINCIPAL CONCLUSION OF THE TEXT IS:"
1080 GOSUB 1280
1090 GOTO 1160
1100 PRINT "THE DEDUCTIVE INFERENCE FROM THE ARGUMENT IS:"
1110 GOSUB 1280
1120 GOTO 1160
1130 PRINT "THE INDUCTIVE INFERENCE FROM THE ARGUMWENIMENT IS:"
1140 GOSUB 1280
1150 REM END OF GRAMMAR PRINT SECTION
1160 PRINT
1170 PRINT "TYPE '1' FOR ANOTHER GRAMMAR ELEMENT, '2' FOR ANOTHER"
1180 PRINT "TEXT, '3' TO PRINT OUT THE ENTIRE TEXT, OR '4' TO END"
1190 PRINT "THE PROGRAM:"
1200 PRINT
1210 INPUT X
1220 PRINT
1230 IF X=1 THEN 650
1240 IF X=2 THEN 90
1250 IF X=3 THEN 1330
1260 IF X=4 THEN 1690
1270 PRINT
1280 REM TO PRINT STATEMENT LINES
1290 PRINT
1300 PRINT "THIS IS A SAMPLE STATEMENT LINE FROM A TEXT"
1310 PRINT
1320 RETURN
1330 REM TO PRINT OUT THE ENTIRE TEXT IF REQUESTED
1340 IF A$=P$ THEN 1380
1350 IF A$=Q$ THEN 1400
1360 IF A$=R$ THEN 1420
1370 IF A$=S$ THEN 1440
1380 GOSUB 1530
```

```
1390  GOTO 1460
1400  GOSUB 1570
1410  GOTO 1460
1420  GOSUB 1610
1430  GOTO 1460
1440  GOSUB 1650
1450  GOTO 1460
1460  REM CHOICE TO END-RUN OR CONTINUE PROCESSING
1470  PRINT
1480  PRINT "TYPE '1' TO END THE PROGRAM OR '2' IF YOU WISH TO CONTINUE
:"
1490  PRINT
1500  INPUT W
1510  IF W=1 THEN 1690
1520  GOTO 90
1530  REM TEXT FOR P$
1540  PRINT
1550  PRINT "THIS IS A SAMPLE TEXT CONTAINING THE WORD 'COMPILER'"
1560  RETURN
1570  REM TEXT FOR Q$
1580  PRINT
1590  PRINT "THIS IS A SAMPLE TEXT CONTAINING THE WORD 'MARBLES'"
1600  RETURN
1610  REM TEXT FOR R$
1620  PRINT
1630  PRINT "THIS IS A SAMPLE TEXT CONTAINING THE WORD 'PIGS'"
1640  RETURN
1650  REM TEXT FOR S$
1660  PRINT
1670  PRINT "THIS IS A SAMPLE TEXT CONTAINING THE WORD 'VEHICLES'"
1680  RETURN
1690  REM END OF PROGRAM
1700  PRINT
1710  PRINT "THIS IS THE EMND OF THE IRS SIMULATION PROGRAM."
1720  PRINT "THANK YOU FOR SHOPPING WITH US TODAY."
1730  END
```

LIST OF REFERENCES

ANDERSON, Digby, C. Some organisational features in the local production of a plausible text. IN Philosophy of the Social Sciences, Vol. 8, 1978, pp. 113-135.

ANSCOMBE, G.E.M. An introduction to Wittgenstein's tractatus. London, Hutchinson, 1967, 179p.

ATTAR, R. et al. KEDEMA- Linguistic Tools for Retrieval Systems. IN Journal of the A.C.M., Vol. 25(1), Jan. 1978, pp. 52-66.

AUSTIN, Derek. PRECIS: a manual of concept analysis and subject indexing. London, The British Library, 1974.

BAR-HILLEL, Yehoshua. Language and Information: selected essays on their theory and application. London, Addison-Wesley, 1964, 388p.

BARTLETT, G. The retention of stories in human memory. New York, Elesvier, 1932.

BARTSCH and VENNEMANN(1973)-see p.200.

BEKTAEV, K.B., et al. Engineering Linguistics. IN Linguistics, Vol. 194, 1977, pp. 43-52.

BELKIN, N.J. and DEA, W. Beyond the sentence: clause relations and textual analysis. (To appear in Informatics 3, London, Aslib, 1979.) 18p.

BELKIN, N.J. A concept of information for Information Science. Unpublished Ph.D. thesis, University of London, 1977.

BELKIN, N.J. and ROBERTSON, S.E. Information Science and the phenomenon of information. IN JASIS, July-Aug. 1976, pp. 197-204.

BELKIN, N.J. Internal Knowledge and external information. (A paper presented at 'The Cognitive Viewpoint', SHENT, Mar. 1977(a)) 15p.

BERTSCH, Eberhard. The Programming Language COMSKEE. Revised report of the project in the Department of Linguistics at the University of Saorbruken, 1978.

BOBROW, D.G. and WINOGRAD, T. An overview of KRL, a Knowledge Representation Language. IN Cognitive Science, Vol. 1(1), 1977.

BOOK, Ronald, V. Simple representations of certain classes of languages. IN Journal of the A.C.M., Vol. 25(1), Jan. 1978, pp. 23-31.

BOTTLE, R.T. and PERRY, K.R. Breed and Cultivator Names in Agricultural Literature. IN The Information Scientist, Vol. 11(1), Mar. 1977, pp. 19-23.

BROOKES, B.C. The developing cognitive viewpoint in Information Science. IN Journal of Informatics, Vol. 1(2), 1978, pp. 55-63.

BRUDERER, Herbert, E. Handbuch der maschinellen und marchinenun terstutzen sprach uberstetzung. (English translation in print.) Verlog Dockumentation, Munich, 1978.

BUXTON, A.B. and MEADOWS, A.J. The variation in the information content of titles of research papers, with time and discipline. IN Journal of Documentation, Vol. 33(1), March 1977, pp. 46-52.

CHARNIAK, Eugene. A framed painting: the representation of a common sense knowledge fragment. Working paper 26, University of Geneva, Institute for the Study of Semantics and Cognition, 1976.

CHOMSKY, Noam. Aspects of the theory of syntax. Cambridge Mass., MIT Press, 1965.

CHOMSKY, Noam. Syntactic structures. Manten, The Hague, 1957.

CLEMENT-DAVIES, Ceuan. Reference retrieval by user-negotiated term frequency ordering within a dynamically adjusted notional 'document'. Proceedings of Informatics 4 (1978) - not yet published.

COOPER, W.S. and MARON, M.E. Foundations of probabilistic and utility - theoretic indexing. IN Journal of the A.C.M., Vol. 2(1), Jan. 1978, pp. 67-80.

DEBONS, Anthony (ed). Information Science: search for identity. New York, Marcel Dekker, 1974.

DIJK, Teun A. (1977) - see p.200.

DREIZIN, Felix and SHENHAR, A. The FOCUS Project: folklore computerised studies. Techored Report No. 1. University of Haifa Press, 1978.

FISHER, John. Identifying the performative verb in human discourse. Proceedings of the Aslib Informatics 4 conference, March 1977.

GILBERT, G.N. The transformation of research findings into scientific knowledge. IN Social Studies of Science, Vol. 6, 1976, pp. 281-306.

GOSHAWKE, Walter. Number language, word processing and information retrieval. IN Journal of Informatics, Vol. 3(1), April 1979, pp. 45-49.

HARRIS, Z. Mathematical Structure of Language. New York, Wiley-Interscience, 1968.

HOCKEY, Susan M. (1978) - see p.200.

HOLMES, V.M. and WATSON, I.J. The Role of Surface Order and Surface Deletion in Sentence Perception. IN The Quarterly Journal of Experimental Psychology. Vol. 28(2), May 1976, pp. 155-166.

HUGHES, G.E. and LONDEY, D.G. The elements of formal logic. London, Methuen, 1965.

HUNT, H. and SZYMANSKI, T.G. Lower bounds and reductions between grammar problems. IN Journal of the A.C.M., Vol. 25(1), Jan. 1978, pp. 32-51.

HUTCHINS, W.J. On the problem of 'aboutness' in document analysis. IN Journal of Informatics, Vol. 1, 1977, pp. 17-35.

KAY, Martin. Experiments with a powerful parser. Santa Monica, Rand Corporation, 1967.

KEMPSON, Ruth, M. Presupposition and the delimitation of semantics. C.U.P., Cambridge Studies in Linguistics Series, 1975. 235p.

KITTREDGE, Richard. Textual cohesion within sublanguages: implications for automatic analysis and synthesis. IN Proceedings of the Seventh International Conference on Computational Linguistics, Bergen, Norway, Aug. 17, 1978.

KLEIN, S. Automatic paraphrasing in essay format. IN Mechanical Translation, Vol. 8(3), 1962.

KNOWLES, Francis, E. Recent Soviet Work on computer techniques for representing natural language meaning. IN Proceedings of Informatics 5, Oxford, March 1979 - not yet published (Aslib).

KNUTH, Donald, E. The Art of Computer Programming. Vol. 1, Fundamental Algorithms. California, Addison-Wesley, 1971.

KNUTH, Donald, E. On the translation of languages from left to right. IN Information and Control, Vol. 8, 1965, pp. 607-639.

KUHN, T.S. The Structure of Scientific Revolutions (2nd ed.), Chicago, University of Chicago Press, 1970.

LANCASTER, Frederick, W. Information Retrieval systems: Characteristics, Testing, and Evaluation. New York, Wiley, 1968.

LANSFORD and HOLMES (1979) - see p. 200.

LANGRIDGE, Derek. (A review). PRECIS: a manual of concept analysis and subject indexing by Derek Austin, 1974. IN Journal of Librarianship, Vol. 8(3), July 1976, pp. 210-212.

LAURIERE, Jean-Louis. A language and a program for stating and solving combinational problems. IN Journal of Artificial Intelligence, Vol. 10(1), Feb. 1978, pp. 29-127.

LEACH, E. Culture and Communication: the logic by which symbols are connected. Cambridge University Press, 1976.

LEECH, Geoffrey. Semantics. Pelican, 1974. 386p.

LEVIN, James and MOORE, James. Dialogue Games: meta-communication structure for natural language interaction. IN Cognitive Science, Vol. 1(4), Oct. 1977.

LORD, Robert (1974) - see p. 200.

LOU, S.C. CULT (Chinese University Language Translator), FBIS Translator on M.T., 1976 , IN American Journal of Computational Linguistics, 1976.

McCAWLEY, J.D. (1971) - see p. 200.

MANDLER, J.M. and JOHNSON, N.S. Remembrance of Things Parsed. IN Cognitive Psychology, Vol. 9, 1977, pp. 111-151.

MARTIN, James. Computer Data-Base Organisation (2nd ed.), N.J., Prentice-Hall, 1977.

MARTYN, John. Guest Editorial. IN The Information Scientist, Vol. 12(3), Sept. 1978, pp. 81-82.

MOSS, Chris. Chinese language becomes a bit faster. IN New Scientist, Vol. 77(1090), Feb. 1978, pp. 418-420.

NORMAN, Donald, A. and RUMELHART, David E. (eds). Exploration in Cognition. San Francisco, Freeman, 1975, 430p.

ORAN, U. (1978) - see p. 200.

PROPP, V. Morphology of the folk-tale. Austin, University of Texas Press, 1968.

RANGANATHAN, S.R. Prolegomena to library classification (2nd ed.), London, The Library Association, 1957.

RUMELHART, David. Notes on a schema for stories. IN D. Bobrow and A. Collins (eds.), Representation and understanding : studies in Cognitive Science. New York, Academic, 1975.

SAGER, Naomi. Sublanguage Grammars in Science Information Processing. IN JASIS, Jan-Feb., 1975, 10-16.

SAGER, Naomi. Syntactic analysis and natural language. IN Advances in computers, Vol. 8, pp. 153-187, 1967.

SALLIS, Philip J. Concept parsing rules for generating an information structure from science text. IN Journal of Informatics, Vol. 2(2), Aug. 1978, pp. 107-116.

SALLIS, Philip J. A partial-parsing Algorithm for Natural Language Text Using a Simple Grammar for Arguments. IN ALLC Bulletin, Vol. 6, 1978(a)pp. 170-176.

SALLIS, Philip J. Text processing: a matter of definition or application. IN Program, Vol. 12(4), 1978(b)pp. 185-187.

SALTON, G. Experiments in multi-lingual information retrieval. IN Information Processing Letters, 2, 1973, 6-11.

SALTON, G., YANG, CS. and YU, C.A theory of term importance in automatic text analysis. IN JASIS, Vol. 26(1), 1975, pp. 33-44.

SCHANK, Roger C. and COLBY, Kenneth M. (eds.) Computer models of thought and language. San Francisco, Freeman, 1973, 454p.

SCHANK, R.C. Conceptual information processing..North-Holland, 1975. See also Schank, R.C. and the Yale A.I. Project Group. SAM - A story understander. Yale University Computer Science Research Report 43, 1975.

SCHANK, R.C. The Structure of Episodes in Memory. IN D.G. Bebrow and A.M. Collins (eds.), Representation and Understanding: Studies in Cognitive Science. New York, Academic Press, 1975, pp. 237-272.

SHANNON, C.E. A Mathematical Theory of Communication. IN Bell System Journal, Vol. 27, 1948, pp. 379-423.

SHANNON, C.E. and WEAVER, W. The Mathematical Theory of Communication. Chicago, University of Illinois Press, 1949.

SHREIDER, Yu. A. Informatsiia i meta-informatsiia (Information and meta-information). IN Naucho-Tecknich-eskaya Informatsiya, series 2, No. 4, pp. 3-10, 1974. English translation available in Automatic documentation and mathematical linguistics, Vol. 8(2), 1974.

SMALL, Henry G. Cited documents as concept symbols. IN Social Studies of Science, Vol. 8, 1978, pp. 327-340.

SPARCK JONES, Karen and KAY, Martin. Linguistics and Information Science. New York, Academic Press, 1973.

SPARK JONES, Karen and KAY, Martin. Linguistics and Information Science: a postscript. IN Natural Language in Information Science by Walker, D.E., Karlgen, H. and Kay, M. (ēds.). Stockholm, Skriptor, 1977, pp. 183-192.

SPARK JONES, Karen. Performance averaging for recall and precision. IN Journal of Informatics, Vol. 2(2), Aug. 1978, pp. 95-106.

SPIESAL-ROSING, Ina. Bibliometric and content analysis. IN Social Studies of Science, Vol. 7, 1977, pp. 97-113.

THORNDYKE, Perry W. Cognitive structures in comprehension and memory of narrative discourse. IN Cognitive Psychology, Vol. 9, 1977, pp. 77-110.

TOCATLIAN, J.J. Are Titles of Chemical Papers becoming more informative. IN JASIS, Vol. 21, 1970, pp. 345-350.

TOMA, Peter P. SYSTRAN as a multi-lingual machine translation system. IN Proceedings of Commission of the European Communities Third European Congress on Information Systems and Networks, Luxembourg, Vol. 1, 3-6 May 1977, pp..569-581.

TRAVERS, Robert M.W. Man's Information System. Pennsylvania, Chandler, 1970, 175p.

VANDIJK, Teun A. Complex Semantic Information Processing. IN Natural Language in Information Science, by Walker, D.E., Kowlgen, H., Kay, M. (eds.). Stockholm, Skriptor, 1977, pp. 127-167. SEE DIJK, Teun A. van.

WEINGARTEN, Frederick W. Translation of computer languages. California, Holden-Day, 1973, 180p.

WERLICH, Egon. (1976) - see p. 200.

WILKS, Yorick. An Artificial Intelligence approach to machine translation. IN R. Shank and K. Colby, Computer models of thought and language. San Francisco, Freeman, 1973, pp. 114-151.

WILKS, Yorick. Frames for machine translation. IN New Scientist, Dec. 1977, pp. 802-803.

WILKS, Yorick. Frames, Scripts, Stories and Fantasies. A paper delivered at Informatics 4, Annual Conference of the Aslib Co-ordinate Indexing Group, Lancaster, 1976.

WILKS, Yorick. Time flies like an arrow: the analysis of ambiguous phrases. IN New Scientist, Dec. 1977(a), pp. 696-698.

WINOGRAD, Terry. A Procedural Model of Language Understanding. IN R. Shank and K. Colby, Computer Models of Thought and Language. San Francisco, Freeman, 1973, pp. 152-186.

WOLFF, J.G. The discovery of segments in natural language. IN The British Journal of Psychology, Vol. 8(1), Feb. 1977, 97-106.

YOVITS, M.C. Information Science: Toward the Development of a True Scientific Discipline. IN JASIS, Vol 20, 1969, pp. 369-376.

REFERENCES ADDENDA

BARTSCH, R. and VENNEMANN, T. Semantic structures: a study in the relation between semantics and syntax. Frankfurt, Athenäum Verlag, 1973.

DIJK, Teun A. Van. Complex Semantic Information Processing. IN Natural Language In Information Science, by D.E. Walker, H. Kowlgen and M. Kay (eds). Stockholm, Skriptor, 1977, pp: 127-167.

HOCKEY, Susan M. Colloquium on the use of computers in textual criticism: a report. IN Bulletin of the Association for Literary and Linguistic Computing, Vol.6(2), 1978, pp: 180-182.

LANGFORD, J. and HOLMES, V.M. Syntactic presupposition in sentence comprehension. IN Cognition, Vol.7, 1979, pp: 363-383.

LORD, Robert. Comparative Linguistics. E.U.P., 1974.

McCAWLEY, James D. Where do noun phrases come from? IN D.D. Steinberg and L.A. Jakobovits (eds.), Semantics: an interdisciplinary reader in philosophy, linguistics and psychology. C.U.P., 1971.

NIDA, Eugene A. Componential analysis of meaning. Menton, The Hague, 1975.

ORNAN, U. Generating and transforming by a computer without a dictionary. IN Bulletin of the Association for Literary Computing, Vol.6(3), 1978, pp: 280-291.

WERLICH, Egon. A text grammar of English. Heidelberg, Quelle and Meyer, 1976.