[From*: Journal of Informatics,* vol.1, no.1, April1977,pp.17-35]

ON THE PROBLEM OF 'ABOUTNESS' IN DOCUMENT ANALYSIS

W. John Hutchins
(The Library, University of East Anglia, Norwich)

Summary

One of the most crucial problem areas of information science concerns the identification of what documents are 'about'. This paper seeks to define the notion of 'aboutness ' within the context of recent work in text linguistics. It describes, first, the essential communicational structures of sentences, paragraphs and texts in terms of theme, rheme and thematic progression, connectors of clauses and sentences, and semantic progression. It then identifies the basic features of the global structures of narrative and expository texts, describes the interaction of macro- and micro- structure in the interpretation of texts and the role of presupposed 'states of knowledge ' in both text production and text comprehension. Finally, it is argued that for the purposes of information systems the 'aboutness' of documents is to be found among the presuppositions of authors concerning the knowledge of their potential readers.

1. INTRODUCTION

The analysis of the content of documents is probably one of the most important activities of any information system. Finding out what documents are about and summarising their contents are the primary functions of abstractors and indexers of all kinds, whether they work for multinational abstracting services, for national bibliographies, for university or public libraries, for specialised commercial or industrial information services, or for any body providing information about published records. Yet it is true to say that this most crucial component of the activities studied by information science has been greatly neglected. There is indeed a very common attitude among information scientists that we do not need to know how indexers arrive at a particular description of the contents of a document; all that matters is whether it enables users to find the document when required. We could probably feel happier with this view if we were not all painfully aware of the inadequacies of the abstracts and indexes which are actually produced. An understanding of the processes of indexing and abstracting will not, of course, lead automatically to any improvement in the quality of indexes and abstracts; but it is certainly arguable that if information science is to make any genuine theoretical advances it must seek to understand this most central activity of any information system.

From a broader perspective there is little doubt that summarisation of some kind is performed by all readers every time they read a document or any kind of text. The ability to say what a text is about must be regarded as one facet of our ability to understand a text; if we do not understand a text we find it difficult to say what it is about. It is therefore somewhat unfortunate that summarisation has been neglected by linguists as much as it has by information scientists. It is only now with renewed interest in the linguistic structures of text that we can find even tentative suggestions about what takes place (cf. Dijk 1972). The present essay can offer no more than the outlines of a linguistics of the processes of summarisation and of how indexers decide the aboutness of documents.

2. THE CONTENT OF DOCUMENTS

The first question to be asked is 'what is meant by the topic of a document in the context of an information system?'

On the surface the answer to our question would seem to be simple: the topic of a document is the subject description on an index entry relating to that document. But in fact there is rarely a straight equation of subject description and 'what the document is about', e.g. in postco-ordinate indexing systems the subject description on an index entry may represent only part of the document's content. The resources of the documentary language used in the system may itself preclude the formulation of subject descriptions expressing the whole topic as seen by the indexer. For these and other reasons that we cannot elaborate here (cf. Hutchins, 1975), we must conclude that the subject description is merely one form of expression of some part of what the document is about. By contrast, the *topic* of a document should be regarded as the summarisation of its content for the purposes of an information system, irrespective of the documentary language in which it may be expressed.

What do we mean by the content of a document? To answer this we need to be clear about the distinction between the 'sense' of a linguistic expression and the 'reference' of that expression.

A particular word, for example *father,* may be used to talk about many different persons. Some of these individuals may be referred to by other words, such as *policeman, bricklayer, carpenter* or *doctor,* or by longer expressions such as *the man standing by the window.* The choice of a particular expression to refer to a particular individual is determined by the appropriateness of the expression on the occasion, and whether an expression is appropriate or not depends primarily on its meaning or 'sense'. The sense of an expression is determined principally by its relationship to other expressions of the language. For example, *father* is related as a kinship term to other words such as *mother, son, daughter,* etc. and it is also related to more generic terms such as *man, male, human, animate,* etc. These relationships determine its sense. They specify, for example, that if any individual is to be appropriately referred to as a *father* it must have the properties of 'humanness', 'maleness', 'adulthood', etc. In other words, the sense of a word determines the range of its potential referents. In isolation a word has a sense, but it has no actual referent; it can have a referent only in a particular textual context.

Similarly for sentences. We may say that in isolation a sentence has a sense, since it conveys some meaning to potential readers or hearers, but that it has no reference. Only when uttered on a particular occasion in a specific context does a sentence have a reference. If I say *The big house on the hill has been bought by a millionaire* then clearly some meaning will be conveyed to every speaker of English. But without knowing the context in which it is spoken nobody can know the specific building or the specific person being referred to.

We may draw a similar distinction between the sense of a text and the reference of a text. In order to understand a text a reader does not need to know exactly which particular referents the author may have had in mind when writing the text.

If, however, we are concerned with the truth or falsity of what is being said or written, then knowledge of the senses of expressions is not sufficient. To take a familiar example from Bertrand Russell, we cannot say whether the statement *The King of France is bald* is true or false outside a particular referential situation. As applied to 20th century France, we can find no reference for *the King of France;* the statement is inappropriate. As applied to 18th century France, however, there is a referent, and we are able to test the truth or falsity of the statement. Thus, while the sense remains constant, the reference varies – and so too, in consequence, does the truth-value.

Similarly for texts. The truthfulness *of* an author cannot be ascertained from what he says alone; it can be determined only by testing his statements in appropriate referential situations. Indexers are not concerned with the truth-value of documents, nor with the particular images in the minds of authors (cf. Fairthorne, 1961). They are concerned solely with the sense of texts - this is what is meant by the content of documents.

In this way, we see that a document has a 'sense' that is independent of its author and, we must add, of any of its readers. As such it attains an autonomous existence as part of what Karl Popper has called 'World 3', the world of objective knowledge (Popper 1972). As such it has properties and relationships that its author may not know of – and perhaps might dispute. What this means in practical terms is that the indexer need not be concerned with – perhaps never should be concerned with - what the author himself thinks of as the topic of his text and its relationships to other texts.

What it does not mean, however, is that the sense of a document can be 'discovered' in a pure, abstract, unadulterated state. Every reader interprets a text according to his own knowledge and environment; every reader has his own idea of what the 'topic' may be. The indexer's task is to take as broad a view as possible of what others may 'read into' a text.

3. THEME, RHEME AND THEMATIC PROGRESSION

In order to tackle the linguistic problems of summarisation we need to have a satisfactory account of text structure.[*] One essential component of text structure is the way in which it reflects a progressive accumulation of semantic information, how one segment of a text builds upon earlier parts, how an author can start from something known to his readers in order to communicate something not yet known. To understand the mechanisms of text progression (as we may call this component of text structure) we must first consider how the communicative dynamism of a message influences the syntactic structure of sentences.

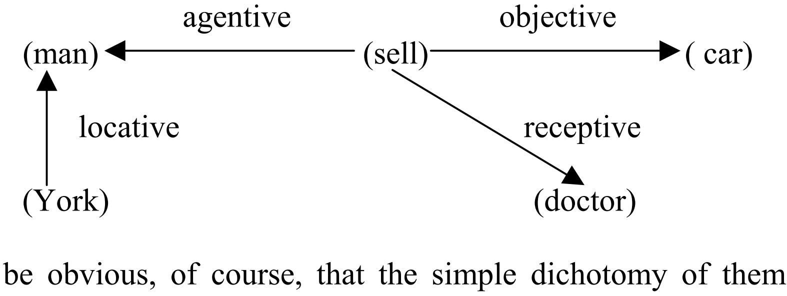
From the point of view of its communicational role we may say that a sentence has two basic parts, a theme and a rheme. The theme represents elements that are related in some ways to the preceding text or to features of the environment in which discourse takes place. The rheme expresses information which is in some sense 'new' to the hearer or reader or which is otherwise unpredictable from what has been said or written already. In rather crude terms we may say that the theme states what the speaker or writer is going to talk about in that sentence, and the rheme expresses what he wishes to say about it. The theme provides the speaker with a point of departure for what he wants to say. Typically, then, the thematic elements of a sentence are either bound textually to preceding text or assumed as 'given' within the context of the utterance. In surface form such contextually and textually bound elements are realised by the use of such anaphoric devices as pronouns and definite articles, e.g. *the man* or *he* referring to a previously mentioned *man.* In addition, anaphora can be indicated by the use of some more generic expression; for example, rather than repeat the specific noun *policeman* a speaker might refer to *the man.* Such a usage then also permits the addition of further specific information, e.g. *the man in blue standing at the pedestrian crossing.* Anaphora can also be indicated by a partitive expression. After mentioning *a car,* we may refer to its parts as *the engine, the wheels,* etc. The basic function of anaphora is thus to alert the addressee or reader to a repeated reference to some object, concept or phenomenon that has been mentioned before or that can be taken as 'given' from the context.

In the typical sentence, elements of the theme will precede elements of the rheme. To illustrate briefly, consider the following sentences:

(1) (a) The man from York sold a car to a doctor
(b) The doctor was sold a car by a man from York
(c) The car was sold to a doctor by a man from York

In semantic content each sentence expresses the same transaction involving the same participants, e.g. represented by a semantic structure such as figure (2); but in the first sentence only the 'man from York' is textually bound, in the second only the 'doctor', and in the third only the 'car'. A specific feature of English is that where possible the initial element of a sentence is made grammatical subject.

(2)



It should be obvious, of course, that the simple dichotomy of theme and rheme cannot be sufficient to explain all of word order in sentences (Firbas 1966, 1974; Sgall 1974). One factor that is

[*] The following sketch of text structure is based on the work of many authors, of whom probably the most influential have been Dressler (1970, 1972), Dijk (1972), Koch (1973) and Petöfi (1973).

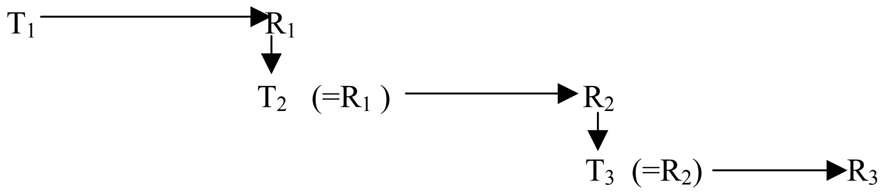
also involved seems to be a ranking of case or role categories, such as Agent, Patient, Recipient, etc., according to order of normal communicative value. In general an Agent has less communicative importance than an action or the effect or result of that action, a Patient or Factitive. Where both Agent and Patient are textually bound or where both are 'new' elements, then the Agent generally precedes the Patient. Hence we normally prefer the active sentence *A girl broke a vase* to its passive 'equivalent' *A vase was broken by a girl*; and this normal preference is indicated by the 'markedness' of the passive verb form.

Among elements of the rheme this seems to be the major ordering principle concerned - but to demonstrate this adequately would take us far beyond the aims of this paper. For elements of the theme, however, other more important factors are also involved. Firstly, we may note that thematic elements expressing the temporal or locative 'setting' of an utterance have less communicative content than those thematic elements expressing the major participants of the event or process being described. They are more peripheral in the sense that they serve often merely as points of orientation for the hearer or reader, placing the event in the general context while adding no fresh information. Typically, such expressions occur before other thematic elements, at the very beginning, e.g. *The next day the doctor was sold a car.* Not all temporal or locative phrases have this text function; in many cases they convey 'new' information and thus appear in the rheme: *The doctor was sold the car on Thursday.*

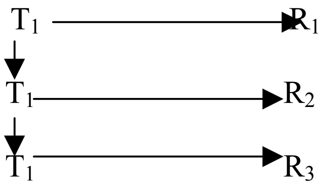
As for the remaining thematic elements, those expressing part of the propositional nucleus, one of them is selected to come before the others, and in English generally made grammatical subject. The selection is based on one or two criteria. An element is favoured either if it has occurred in this position in the preceding sentences or if it is the most recently mentioned of the thematic elements. Such an element may thus be related to the preceding sentence in one of two ways; either it repeats (anaphorically or generically or partitively) part of the theme or it refers again to some element of the rheme.

We have, therefore, two basic types of sentence progression from the viewpoint of theme-rheme articulation or 'thematic progression' as we shall call it, following Daneš (1974): linear progression, where the favoured thematic elements relate to elements of a preceding rheme (figure 3); and parallel progression, where theme remains constant (figure 4).

(3)



(4)



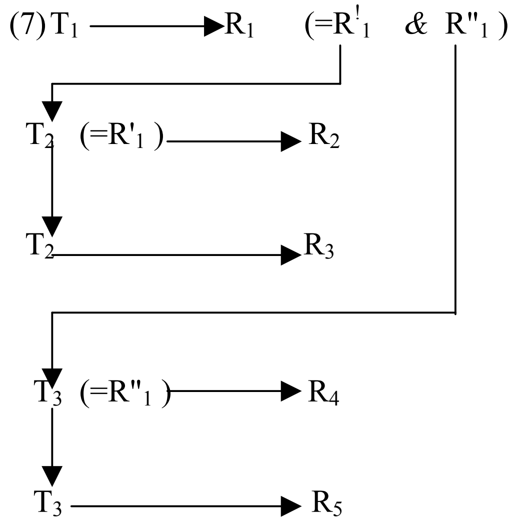
Linear progression may be illustrated by a sentence sequence such as:

(5) The boy was reading a book. It was about armadillos. They are found in South America.

An equally banal illustration for parallel progression could be:

(6) The boy was reading a book. He had been given it for his birthday. He was ten years old.

In the majority of cases, however, we find a mixture of the two types. A common example is the exposition of a split rheme by parallel progression:



This may be illustrated by a paragraph such as:

(8) All substances can be divided into two classes: elementary substances and compounds. An elementary substance is a substance which ...(etc). A compound is a substance which ...(etc).

For each type of thematic progression, we see that the first sentence provides the starting point or foundation for the following sentences. In this sense it may be regarded as a whole as the theme of the paragraph, where the subsequent sentences contribute the rheme. Such an observation is by no means new. Christensen (1967) for example, refers to the first sentence of a paragraph as the 'topic sentence'. Subsequent sentences are related to it either co-ordinately or subordinatively; they add qualifications, attributes, details, or make comparisons concerning elements of the first sentence.

We are thus lead to some tentative formulations about what is meant by the topic of a paragraph and the topic of a sentence. In rough terms we may say that in most instances a paragraph is 'about' whatever is communicated in the first sentence of its thematic progression. As for sentences, it would seem in general that what we usually mean by the topic is the thematic element forming the major link to the immediately preceding sentence, i.e. the element referred to earlier as the favoured thematic item, the one generally selected in English as grammatical subject.

We have discussed in this section only the most normal and most typical cases. In practice, matters are far more complex. Firstly, there is no straight correspondence between anaphora and the use of definite articles or pronouns: a definite article does not necessarily indicate that an anaphoric relation is present, nor is anaphora necessarily realised by a definite or pronominal form. Secondly, it is by no means uncommon for the beginning of sentences to be occupied by elements of the rheme. The reasons are usually related to matters of emphasis and style. A common device is the use of an anticipatory *it,* as in: *It was a doctor that was sold the car by a man from York.* Here the rhematic element *doctor* has been brought forward perhaps to correct some misunderstanding on the part of the addressee. A similar effect is achieved by simply stressing a particular element, without changing the word order, e.g.

(9) (a) *Harry* sold the car to Bill
(b) Harry *sold* the car to Bill
(etc)

Lastly, the actual form of thematic progression in texts is rendered more complex by the close interaction of other aspects of text progression, chiefly features of semantic progression. We have already mentioned some components of semantic progression when discussing anaphoric relations. Whenever a particular object, concept or phenomenon is referred to again in a text not by the use of the same expression but by the use of some more specific description, then the speaker is contributing further information about the object. He is furthering the communication, and the 'new' information is assisting the semantic progression of the text. Another aspect of semantic progression was touched upon with reference to paragraph structure. We spoke of the sentences following the 'topic' sentence as adding details, qualifications or comparisons, i.e. as elaborating, expanding and illuminating the message being conveyed.

4. CONNECTORS AND SEMANTIC PROGRESSION

From the viewpoint of text structure, a major role in semantic progression is played by 'sentence connectors'. So far we have discussed the communicational structure only of simple sentences, that is to say: sentences composed of only one clause with a single finite verb form. Clauses may be defined as the realisations of single propositions, consisting of an action or process and its participants (the nucleus, e.g. as represented in figure (2)) and its temporal and local setting or environment. It is the function of the connectors to join clauses together into complex sentences and to relate sentences (whether simple or complex) to each other in texts. Connectors of the first kind are the conjunctions, both co-ordinative (e.g. *and, but)* and subordinative (e.g. *before, because, if).* Connectors of the second kind are the 'sentence adverbs' such as *however, nevertheless,* etc. and such prepositional phrases as *before this, after that.* Semantically, however, the connectors form *a* coherent group with equivalent functions, namely the joining of clauses into semantic progressions where each clause has its own theme-rheme articulation and its place in a thematic progression.

There have been many attempts at classifying conjunctions, prepositions, adverbs and other connectors, and many thorough analyses of their semantic content and inter-relationships. For our purposes we may be satisfied by a broad division into five main classes according to the type of progression they manifest. (This classification comes from Longacre(l970)).

Firstly, there are the various temporal connectors, to mark the chronological progression of a narrative. Corresponding to the subordinating conjunctions *before* and *after* we have the sentence adverbs *afterwards, beforehand* and *then:*

(10) (a) After Bill arrived, Jim left
(b) Bill arrived before Jim left
(c) Bill arrived. Then Jim left
(d) Bill arrived. Afterwards Jim left

(Ignoring slight differences of emphasis the same temporal sequence of events is being reported in each sentence.)

Other temporal expressions are those of concurrent time, e.g. the conjunction *while,* and the adverbial *meanwhile* and *at the same time,* etc. subsequent time, e.g. *until,* and preceding time, e.g. *since.*

A second group of connectors are those expressing teleological relations, for example: purpose, by *in order to* and *in order that,* and cause, by *because* and the sentence adverb *therefore.*

(11) (a) Because Mary had forgotten to buy some bread, Harry went to the shops
(b) Mary had forgotten to buy some bread. Therefore Harry went to the shops

Other connectors in this group are those of circumstance, reason and result.

The connector for circumstance is often *since:*

(12) (a) Since the soup was too hot we could not eat it

The connector for result is the conjunction *so:*

(12) (b) The soup was too hot, so we could not eat it.

Both circumstance and result may be regarded as weaker forms of the causative, thus the replacement of *since* and *so* by *because* or *therefore* is possible on many occasions, as it is here. Indeed the result connector is frequently so weak in semantic force that it can be omitted altogether, giving a simple sequence of sentences with no overt connector:

(12) (c) The soup was too hot. We could not eat it.

Similar remarks apply to the reason connector, for example, the conjunction *for:*

(13) (a) I gave into his demands, for there was nothing else to do.

and:

(b) I gave into his demands. There was nothing else to do.

or intensified as a causative construction:

(13) (c) I gave into his demands because there was nothing else to do.

We may note also the inverse relationship between the result and reason connectors. A result connector *so* can be replaced by a reason connector *for* with a straight transposition of the clauses:

(14) (a) The soup was very hot, so we could not eat it.

(b) We could not eat the soup, for it was very hot.

and vice versa. Consequently when we encounter consecutive sentences with no explicit sentence connector we may be dealing with either a progression of result or of reason, and only the semantic content can determine which is the case in specific instances:

(15) (a) I gave into his demands. There was nothing else to do.

(b) There was nothing else to do. I gave into his demands.

A third group of connectors includes those expressing such logical relations as condition, concession, contrafactualness and correlation. Concession may be expressed by such clause connectors as *although, even though,* and by sentence adverbs *nevertheless* and *yet,* and the prepositional *in spite of,* for example:

(16) (a) We advised him not to go. Nevertheless he went.

(b) Although we advised him not to go, he went.

(c) In spite of our advice, he went.

For the conditionals we have the familiar *if...then* construction, which however has no corresponding sentence connector. The contrafactual relation is a conditional restricted to past time:

(17) If Jim had not worked overtime, he would have arrived by seven

and the correlative relation is a two-way conditional, expressed by *as...so:*

(18) As Maine goes so goes the nation

A fourth group of connectors may be called those of concatenation: co-ordination, expressed typically by the conjunction *and* as either clause or sentence connector; alternation, expressed by the conjunction *or,* intensified as needed by *either* or by an adverb such as *alternatively;* and lastly, the various connectors expressed by *but,* of which three main ones may be identified. Antithesis, e.g.

(19) (a) He is not a paragon of virtue, but he is a good man

Typical is the thematic progression with constant theme, permitting the contraction into a single clause sentence in which the connector joins two rheme elements:

(19) (b) He is not dead, but alive

The antithesis may be further marked by the adverbial *on the contrary,* which may also on occasion stand alone in place of *but:*

(19) (c) He is not dead. On the contrary, he is alive.

Secondly, contrast – e.g.

(20) My horse is black, but yours is white

where *but* may be replaced or emphasised by the adverbial *by contrast.* Finally, *but* may occur where an expected consequence is denied:

(21) (a) They set out for Paris, but they did not arrive

In this sense, the corresponding sentence adverb is *however:*

(21) (b) They set out for Paris. They did not arrive, however.

It should also be noted that connectors of concatenation are often employed to underline the thematic progression of a text. For example, the introduction of the themes in the development of a split rheme (fig. (7)) above may be signalled by *alternatively* or *on the other hand* if the relation between the sections is one of alternation, and by *however, on the contrary, by contrast,* etc. if it is one of antithesis, contrast, etc. Progressions with constant theme may be made more explicit by such sentence adverbials as *furthermore, in addition, also, too,* or even more explicitly by the numbering of points: *first, secondly, thirdly,* etc.

This brings us to the last group of connectors, those of paraphrase, recapitulation and illustration. Typical expressions of these connectors are *in other words* and *in brief* for paraphrase, *I say again* and *to repeat* for recapitulation, and *for example* for illustration.

To conclude this brief consideration of clause and sentence connectors we must emphasise one point. There is no necessary relationship between the semantic progression of a text and the logical or chronological succession of the argument or of the narrative. We see this most clearly with the temporal connectors. Although an event X may have 'really' happened before an event Y, a speaker may elect to say *Before Y there was X* or *There was Y after X.* Similarly for logical sequences: instead of putting a conditional before its consequent, a speaker may choose to invert them: to say not *If X then Y* but *Y if X.* Such inversions may also at times involve larger segments of text; for example, the 'topic sentence' of a thematic progression might be placed at the end of a paragraph, thus serving as a kind of summary introduced perhaps by *in other words.* This could happen in the mixed type of thematic progression with a split rheme (fig. 7).

5. GLOBAL TEXT STRUCTURE

In themselves, thematic progression and semantic progression tell us nothing about the topic of a text as a whole. Fairthorne (1969) has drawn a distinction between the extentional aboutness of a text and its intentional aboutness. The former is defined by the topics of component parts of a text, the topics of its paragraphs, sections, chapters, etc. The latter is the topic of the text as a whole, representing something more than the topics of its parts.

Thematic and semantic progression express only one type of text relationship, namely that of succession. Speakers and writers have no option but to convey their message in linear form; they are constrained by the mental limitations of their audiences. They must proceed gradually, building from the familiar to the novel. A text represents a progressive modification and accumulation of information. Thematic and semantic progression are necessary characteristics of text; they must be present if a sequence of sentences is to be a coherent text. But they do not themselves express the total message of the text, and it is this, the semantic content (the sense) of the text, that conveys the information with which the speaker or writer hopes to modify the state of knowledge of his addressee or reader.

In considering how the total information content of a text is built up from its constituent propositions we must turn to structural properties of a more global nature. Specifically we must consider how a sequence of sentences constitutes an episode of a narrative or a stage of an argument, and how episodes and argument stages are related to each other in a coherent text.

Much of the work on text structures has been concerned primarily with narrative texts, with texts that tell a story, whether true or fictional. But in information systems we are concerned mainly with texts of an expository nature, with texts that describe a state of affairs, that put forward an argument, a theory, or that discuss alternative explanations of some phenomenon. Nevertheless, some analogies between narrative and expository texts seem to be legitimate.

The stimulus for much of the work on narrative text structure has been the now well known pioneer work of Vladimir Propp (1968) on Russian folk tales. Propp demonstrated that the Russian folk tale could be analysed into an invariant sequence of episodes. These episodes are defined as 'functions' of the principal participants of the tale, e.g. the treachery of the villain, the rescue of the victim by the hero, and so forth. In any particular tale not all the functions need be present, but those which do always occur in a fixed order.

A somewhat similar analysis of narrative structure is to be seen in the more abstract (but perhaps more familiar) characterisation of a story as consisting of a sequence such as Aperture, Setting, Inciting Moment, Developing Conflict, Climax, Denouement, Final Suspense, Closure (Longacre 1974). The Aperture is one of the conventional openings of stories, e.g. *'Once upon a time……‘* and may often be absent, particularly in written literature. The Setting consists of those parts introducing the main characters and the principle location. In the Inciting Moment some event, object or person is introduced that disturbs the particular situation described in the Setting. In the Developing Conflict section the disturbance becomes more critical, problems and complications become more involved. The crisis intensifies until everything comes to a head in the Climax. Something then happens in the Denouement that makes possible a resolution of the conflict, a way out can be seen. In the Final Suspense the outcome is kept in some doubt by fresh complications; but everything is brought finally to a happy or unhappy end in the Closure.

Although clearly not all narratives conform to this pattern, it is nevertheless a common and readily identifiable type. When we turn to expository texts a common pattern is less easily recognised. One reason for this must surely be that the semantic progressions of expository texts can be of so many different kinds. In narrative texts the connectors are primarily those of time, but in expository text we encounter the whole variety of logical, teleological and other non-temporal connectors. Nevertheless, we can perhaps identify one common type of text, that frequently found in scientific papers. We may represent it as follows:

(22)

the ‘problem’ {
statement of 'current' hypothesis
tests of hypothesis
disproof of hypothesis
statement of 'problem'

the 'solution' {
statement of 'new' hypothesis
tests of hypothesis
'proof of hypothesis
statement of 'solution'

implications of 'solution'

The two basic components are the statement of a 'problem' and a proposed 'solution'. The scientific paper is thus seen as being an integral part of 'normal' science, as defined by Kuhn (1970). A problem has arisen in the interpretation of certain data within an accepted paradigm. The objective of the author is to offer a solution or to suggest where one may be found. The author begins by outlining the current approach to the particular state of affairs in the scientific area with which he is concerned. Then, he may review the evidence that indicates some inadequacy of the current approach; and he may conclude with a demonstration that there is a problem to be solved. In his 'solution' the author may first say in which direction he thinks opinion should change. He provides then some evidence to support his contention and to show that the problem can be solved in this way. Finally he may offer some comments on the implications of his proposal in other problem areas.

Not all sections need be present, as with Propp's functions in Russian folk tales. Very often no tests of the 'current' approach are given, either because its inadequacies are presumed to be widely known already or because the author refers his readers to other texts where such tests can be found. Similarly, the author may offer no testing of his own 'solution'; he may present it 'simply as a programme for future research. On other occasions, he may intend only to demonstrate the inadequacies of the current model without putting forward any alternative solutions; his aim is simply to show there is a problem requiring a solution.

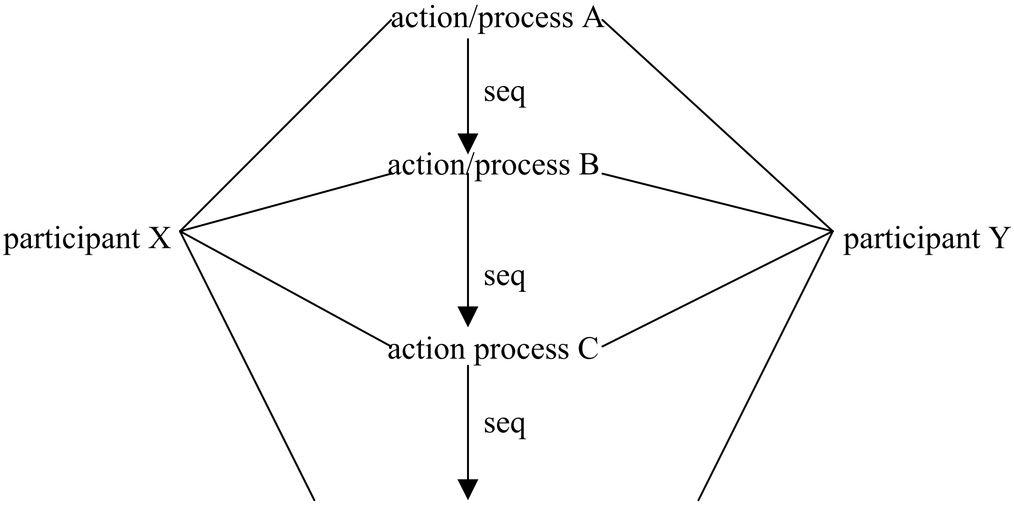
6. MACRO-STRUCTURE AND MICRO-STRUCTURE

The question now arises: What is the relationship of the episode structure of *a* narrative or the argument stage structure of an expository text to the structural properties of atext as defined by semantic and thematic progression? One answer is that we may regard an episode or a stage as a segment of text displaying a single coherent principle of text progression, i.e. a paragraph as we defined it earlier. In its simplest form an episode might be a sequence of sentences or clauses in which one participant is involved in a temporal succession of activities. The participant recurs in each clause as the theme, and the activities constitute the rheme. We have thus a semantic progression based on connectors of time and *a* thematic progression with constant theme.

From the perspective of the text as a whole, an episode or stage may be regarded as one element in a global text progression. In a narrative one episode follows another in a broad chronological succession. In an expository text one stage may be related to its predecessor as, for example, an effect and its cause. In other words, just as we have semantic connectors between one proposition and another we may have similar connectors between one episode or stage and another. We might summarise a story by saying "First X did this, then he did that, then he met Y" and so forth. And we might summarise an argument by saying: "The view that X is true has been shown to be invalidated because of A, B and C. Therefore it is proposed that Y is true. And if Y is true, then Z follows".

From yet another perspective we may regard an episode or argument stage as a generalisation or summarisation of the semantic content of its constituent propositions. Propp's functions may be seen in this light. In a given sequence of clauses we may find the same two participants recurring in propositions whose predicates have similar semantic content, e.g. in a fight where first one participant strikes the other and then receives a blow, and so on; figuratively:

(23)



The generalisation of such a text progression requires the selection of an appropriate predicate generic in sense to all the action/processes expressed. Similarly, other features of a given text progression may be summarised; instead of the individual details of the location or locations in which the events take place, one might give a more general description of the location of the episode as a whole, e.g. "in a forest"; and whereas each individual proposition may indicate a particular point of time, the summarisation in the episode may express the time of the events more generically. We may thus define an episode in terms similar to that of a proposition, as a semantic representation consisting of a predicate and its arguments together with indications of temporal and local setting.

We may thus envisage the representation of the semantic content of a text as a two-level structure: a semantic network representing its underlying propositions, their inter-relationships and their place in the global semantic progression; and a semantic network representing the propositions of episodes and their relationships to each other and within the text progression. Following Van Dijk (1972) we may call them the micro-structure and the macro-structure respectively. Both are essentially of the same form: networks of propositions related by connectors where individual participants occur as arguments in a number of propositions and their occurrences are inter-related by various anaphoric devices. There are good reasons to suggest that what any reader of a text remembers of its content is something like the semantic network of its macro-structure, i.e. he remembers the sequence of the major episodes in a story and something about the chief participants, or in the case of expository text he remembers the general outline of the argument and the main points for and against a particular proposition. What he rarely remembers are the details of micro-structure, the particular sequence of events in a given episode or the specific progression of a given stage of the argument. Even less clearly remembered are the particular linguistic expressions used by the author, the surface forms of the text - with the obvious exceptions of particularly striking literary texts.

7. TEXT COMPREHENSION

We may, therefore, suggest what the linguistic processes may be in the comprehension of a text. Every sentence is interpreted as a complex of propositions whose elements and relationships are to be integrated into the semantic network that has already been established. Integration operates simultaneously at both the level of micro-structure and at the level of macro-structure. The relationships of micro-structure provide the immediate context for the interpretation of the utterance in question; they supply the information necessary for the recognition and understanding of the textually bound elements, they supply the point of departure for the utterance in a given text progression, and they are the repository of what *is* already known about particular participants of the utterance. In short the micro-structural network establishes the semantic coherence of the text and specifies the function of individual propositions within the total complex. But, at the same time, the interpretation of a segment of text contributes also to the establishment of the macro-structure. Every sentence has a role in the expression of a particular episode of a story or of a stage in an argument. At this level, we may envisage the already established macro-structure as providing the context for the interpretation of a particular episode or argument stage. Just as every proposition is related to its predecessors by anaphoric relations and by connectors in a semantic progression, so too *every* episode *is* bound textually to preceding episodes, since characters and locations recur and are elaborated, and every episode is part of the progression of the plot.

If we envisage the comprehension of a text as the building of a semantic network to represent the content of that text, then we must ask what are the foundations on which the reader builds. Clearly he must build substantially upon his previous knowledge of objects, events and phenomena which are referred to in the text. In the general model we are describing we assume that the reader's state of knowledge may be represented as a complex semantic network. In the reading of a text this network provides the context for its interpretation and may be changed or augmented as a result of the interpretation.

What are the constituents of the semantic network representing a reader's state of knowledge? Firstly, there are the complexes representing the conventional senses of individual lexical items, i.e. representing in toto his knowledge of the vocabulary of the language. Next, there are the associative relations that an individual lexical item may have to other items by virtue of certain properties of their referents or by virtue of their common occurrence together in the physical world or when talked about in discourse. These associations may be shared by the majority of the language community, or they may be held by only a relatively small group of speakers. In the latter case the associations may contribute to a specialised usage of the term within the group. Next come the relationships established on the basis of known properties of the referents of lexical items: some of these constitute 'common knowledge', i.e. knowledge shared by all members of the community, others may be familiar only to particular specialised groups, e.g. 'scientific knowledge'. Finally, there are the relationships that are peculiar to the individual reader only - some of these may concern his knowledge of the sense of a lexical item, others may concern his knowledge of what he believes to be true about the referents of lexical items.

What may an author presume his readers to know already? In some cases he may assume only knowledge of the conventional senses of a given part of the vocabulary and certain aspects of 'common knowledge' about the referents of the lexical items involved. On other occasions he may assume more specialised knowledge of the referents, he may assume acquaintance with a certain specialised terminology. If the reader lacks any of the knowledge presupposed by the author then clearly he will have difficulty in relating what the author says to his own network. The points of contact will be missing; what is taken as 'given' will be in fact something new and unknown.

It may be supposed that any part of a given reader's semantic network which is not presupposed by the author but which falls in the general area of the content of the text is open to change in the course of interpreting the text. However, not all parts will be equally susceptible to change. Some configurations may represent strongly held beliefs of the reader that may be virtually unshakeable by any argument. But even those parts less strongly held will be changed only if the reader is convinced by the author's thesis. Therefore, in expository text, the semantic progression from stage to stage of the argument and within each stage is a vital component of the text's message; both macro- and micro-structure must be logically coherent. Even then, an individual reader may be unconvinced if he cannot follow the argument, fails to recognise certain logical or teleological connectors or misinterprets them.

8. SUMMARISATION AND 'ABOUTNESS'

The individual reader of a document will always be interested primarily in information conveyed that is new to him. He will tend to concentrate his attention on those parts of the text content which contribute substantial additions to his state of knowledge. If asked to say what the document is 'about' he will tend to mention some aspect of this 'new' content.

But for obvious reasons, such a definition of a document's 'aboutness' would not be appropriate for an information system serving the needs of many different readers with many different levels and ranges of knowledge. What is needed is a definition of 'aboutness' which is sufficiently general and constant to satisfy all the users of a given information system. From the discussion of this paper two approaches would seem to meet this requirement; firstly, a definition in terms of the total semantic network of a text, and second, a definition in terms of the knowledge base presupposed by the author.

At first sight, the notion that a statement of 'aboutness' should be in some sense a summary of semantic content seems to be inherently plausible. It also has the merit of suggesting mechanizable processes. In terms of the model of text structure presented here summarisation may be seen as the generalisation and reduction of text micro-structure. Generalisation is the process described earlier for the characterisation of an episode or argument stage by a single proposition. Reduction may be defined as the elimination from a semantic network of those elements and relationships which are inessential and unimportant to the development of the main plot or argument. The judgement of what is essential and important and what is not depends, of course, very much on the purpose for which the summarisation is being made, but nevertheless there is one procedure for the reduction of a semantic network which would appear to have general application. This is to use a measure of the relative strengths of the relations contained in the network. On the whole, the more often a particular item or a particular linkage is mentioned in a text the more characteristic is that item or linkage of the text's content. In general the more frequently occurring participants and activities will tend to constitute elements of the macro-structure.

A number of automatic indexing and abstracting systems are based implicitly on this approach to 'aboutness', in that content is measured by calculations of relative frequencies. Of necessity, most are based on counts of word frequencies, and they are for this reason alone unsatisfactory since the same content can be expressed by an almost infinite variety of surface lexical forms and any particular expression may convey a number of different potential senses. But even allowing for this, frequency counts must be used with caution. At the lexical level, the most frequently occurring words are those of the basic vocabulary and of syntactic functions conveying little informational value. At the semantic level, a good proportion of the semantic network of a text must represent the expression of items and relationships of 'common knowledge' known to all speakers of the language. Although these parts are of negligible importance for characterising document content, they will inevitably be frequent components of any text. Clearly, the summarisation of a document must seek to identify those frequently occurring elements of the network that do not correspond to 'common vocabulary' or 'common knowledge'. But the question remains whether even with this qualification and with further refinements, an approach to 'aboutness' based entirely on summarisation is appropriate for the purpose of document analysis in an information system.

If we examine the anaphoric network of frequently occurring text elements we find that typically the first references will appear in the earlier parts of texts. In the initial sections of a text an author will generally establish the major participants of his narrative or argument. Thus, as we have seen, a story frequently has an early section devoted to the Setting of the narration, introducing the chief characters and the main location. Similarly, in an expository text the first section will normally introduce the fundamental components of the subsequent argument, it will state the main ingredients of the 'problem' to be discussed. Therefore, just as the initial sentence of a paragraph may be regarded as the foundation for the thematic progression manifested by the following sentences, so too may some parts of the initial sections of a text be regarded as the foundation for the thematic progression of the text as a whole.

But there is also another sense in which the early parts of a text may be seen as providing the foundations for the text. We have mentioned on a number of occasions that the author must relate what he has to say to what he presumes his potential readers may already know. Whatever substance these presuppositions may have, it is natural that the author should make such links in the initial sections of the text. It is upon the foundations of what he takes as 'given' and 'known' that he can develop the semantic progression of the text.

It is then in the initial sections of a text that the author establishes points of contact for his readers, either by relating what he has to say to a particular context or environment or by relating it to previous discourse or texts. At the same time he introduces the main components of the semantic network of the text which he shall develop in subsequent sections. The analogy to the notion of 'theme' should now be clear. Just as we may refer to the themes and rhemes of sentences and paragraphs, so we may in broad terms refer to thematic parts and rhematic parts of text. The thematic part of a text expresses what the text is 'about', while its rheme expresses what the author has to say about it. In other words, we contend that the 'aboutness' of a document is to be sought in those initial sections where the author introduces the major components of the macro-structure and establishes points of contact with what he assumes to be the 'states of knowledge' of potential readers. He says first what object, event or phenomenon of the presumed 'known' he is going to deal with, and then proceeds to say something about it.

The author may take as his starting point some aspect of a previous text. What another author has presented as 'new' information may well be taken as 'given'. Every text has its forbears (Fairthorne 1961: 173). When these are acknowledged by citations, as happens in most scientific texts, we have the basis for grouping texts by shared citations ('bibliographic coupling'), since if two texts cite the same earlier text they are likely to make the same presuppositions about the 'state of knowledge' of readers and they are likely (with some reasonable probability) to be on the same topic. There is thus an obvious temptation to extend our theme-rheme analogy further: just as successive sentences reflect patterns of thematic progression (figure 24)), so too may successively published texts reflect a similar pattern of relationships (figure (25)).

(24)

$$\begin{tikzcd}
\text{Sentence}_1 : & T_1 \arrow[r] & R_1 \arrow[d] & \\
\text{Sentence}_2 : & & T_2 \arrow[r] \arrow[d] & R_2 \\
\text{Sentence}_3 : & & T_3 \arrow[r] & R_3
\end{tikzcd}$$

(25)

$$\begin{tikzcd}
\text{Text}_1 \text{ :} & \text{Base}_1 \arrow[r] & \text{New}_1 \arrow[dr] \arrow[dd] & & \\
\text{Text}_2 \text{ :} & & & \text{Base}_2 \arrow[r] & \text{New}_2 \\
\text{Text}_3 \text{ :} & & \text{Base}_3 \arrow[r] & \text{New}_3 &
\end{tikzcd}$$

9. 'ABOUTNESS' AND PRESUPPOSED STATES OF KNOWLEDGE

An advantage of this approach to the notion of 'aboutness' with respect to the analysis of documents in information systems is that it explains in a unified fashion a number of otherwise apparently unrelated observations. First, it gives an explanation in text-linguistic terms for the familiar practice of indexers to search for indications of what a document is 'about' in the preliminary and introductory passages of a book or article, and for their almost invariable success in doing so. But it is a practice not unique to indexers. All readers, it seems, expect to find the essential 'aboutness' of a text in its initial passages (cf. Warr, 1966).

Second, whenever anyone consults an information system in search of a document answering a particular information need, he cannot in the nature of things formulate with any precision what the content of that document should be. He cannot specify what 'new' information should be conveyed in an appropriate document. All that he can do is to formulate his needs in terms of what he knows already, his present 'state of knowledge'. What he is looking for, in effect, is a document that starts from a knowledge base with which he can make points of contact, a document which presupposes a state of knowledge with some affinity to his own. In terms of our model, he is searching for a document having some appropriate part of his semantic 'knowledge' network as its theme. He can specify its 'aboutness' only in this way; he cannot say what its global semantic network should contain. Consequently, whether our definition of document 'aboutness' is a valid reflection of the actual practice of indexers or not, there are strong reasons for arguing that the purpose of indexing should be to indicate the 'aboutness' of documents in terms of what knowledge they presuppose. Only in this way can it be possible to relate what users of the system know already to documents telling them what they do not know.

However, not all users will approach the same document with equivalent 'states of knowledge'. The indexer must decide on some average level valid for the majority of the users of the system concerned. Such decisions are somewhat easier in organisations serving a closed specialised community, and correspondingly much more difficult if the system is accessible to all members of the public. In the latter case, the indexer can probably assume only what we have referred to earlier as 'common knowledge'.

The environment of indexing has thus a considerable influence on the formulation of 'aboutness'. Documents already in a collection can also influence the way an indexer expresses 'aboutness'. Like any other reader, the indexer can only judge the 'newness' of a particular document on the basis of his knowledge of other texts. He seeks to integrate what it has to say within an established framework. For the individual reader this framework is his semantic network of 'knowledge'. For the indexer the framework must be the network of document contents in an existing collection, a framework which necessarily reflects previous decisions about the topics of documents.

Finally, there is a persistent and perhaps inherent conflict between what readers regard as the 'aboutness' of a document and what indexers define as its 'aboutness'. Any reader is primarily interested in the 'new' information of a document, in the additions and changes that are made to his knowledge map as a result of reading the text. What he knows already is naturally of no importance; for him the document's importance lies in what he learns from it, whether it answers a particular information need. But for the indexer, 'aboutness' must be formulated in terms of the familiar, the 'given' framework of recorded knowledge. Perhaps inevitably there can be little agreement on the 'relevance' of a particular document to a given topic.

For the reader a document may be relevant if it contributes in some way to the 'solution' of questions that interest him; for the indexer a document may be relevant if it discusses the topic of interest, whatever the value of its contribution on a particular occasion. For the reader, relevance is a function of his current interests and his personal 'state of knowledge'; for the indexer, relevance is a function of the place of the document in the current 'state of knowledge' as a whole, in Popper's world of autonomous 'objective knowledge'.

To summarise the essential features of this approach to document 'aboutness'. We suggest that for the purposes of information systems a summary of the total semantic content of a document is not what is needed. The primary aim of indexing is to provide readers with points of contact, leading them from what they know to what they wish to learn. In document analysis the most important parts of a document's semantic network are those elements that form the knowledge base upon which the writer builds the 'new' information he tends to convey. The restriction of summarisation to the generalisation and reduction of micro-structure fails to take account of the basic communicational structure of texts as manifest in their features of thematic and semantic progression. It is contended, therefore, that for the purposes of information systems the 'aboutness' of a document may be defined in terms of those parts of its generalised semantic network that relate the document to the context of the assumed 'states of knowledge'.

REFERENCES

CHRISTENSEN, F. (1967) A generative rhetoric of the Paragraph. *In his: Notes towards a new rhetoric: six essays for teachers.* New York, Harper and Row, 1967, pp. 52-81.

DANEŠ, F. (1974) Functional sentence perspective and the organisation of text. *In:* DANEŠ, F. (ed.) *Papers in functional sentence perspective.* The Hague, Mouton, 1974, pp. 106-128.

DIJK, T.A. VAN (1972) Some *aspects of text grammars.* (Janua Linguarum, Ser. Maior, 63) The Hague, Mouton, 1972.

DRESSLER, W. (1970) Textsyntax. *Lingua e Stile* 5, 1970, pp. 191-213.

DRESSLER, W. (1972) *Einführung in die Textlinguistik.* Tübingen, Niemeyer, 1972.

FAIRTHORNE, R.A. (1961) *Towards information retrieval.* London, Butterworths, 1961.

FAIRTHORNE, R.A. (1969) Content analysis, specification and control. *Annual Review of Information Science and Technology 4,* 1969, pp. 73-109.

FIRBAS, J. (1966) On defining the theme in functional sentence analysis. *Travaux Linguistique de Prague 1,* 1966, pp. 267-280.

FIRBAS, J. (1974) Some aspects of the Czechoslovak approach to problems of functional sentence perspective. *In:* DANEŠ, F. (ed.) *Papers in functional sentence perspective.* The Hague, Mouton, 1974, pp. 11-37.

HUTCHINS, W.J. (1975) *Languages of indexing and classification: a linguistic study of structures and functions.* Stevenage, Herts., Peter Peregrinus, 1975.

KOCH, W.A. (1973) *Das Textem: gesammelte Aufsätze zur Semantik des Texts.* Hildesheim, Olms, 1973.

KUHN, T.S. (1970) *The structure of scientific revolutions.* 2nd ed. Chicago, Ill., Chicago Univ.Pr., 1970.

LONGACRE, R.E. (1970) Sentence structure as statement calculus. *Language 46* (1970), pp. 783-815. *Reprinted in:* BREND, R.M. (ed.) *Advances in tagmemics* Amsterdam, North-Holland Pub.Co., 1974, pp. 251-283.

LONGACRE, R.E. (1974) Narrative versus other discourse genre. *In:* BREND, R.M. (ed.) *Advances in tagmemics.* Amsterdam, North-Holland Pub.Co., 1974, pp. 357-376.

PETÖFI, J.S. (1973) Towards an empirically motivated grammatical theory of verbal texts. *In:* PETÖFI, J.S. and RIESER, H. (eds.) *Studies in text grammar* Dordrecht, Reidel, 1973, pp. 205-275.

POPPER, K.R. (1972) *Objective knowledge: an evolutionary approach.* Oxford, Clarendon Press, 1972.

PROPP, V. (1968) *Morphology of the folktale.* 2nd ed. Austin, Texas Univ. Pr., 1968.

SGALL, P. (1974) Focus and contextual boundness. *In:* DAHL, Ö. (ed.), *Topic and comment, contextual boundness and focus.* Hamburg, Buske, 1974, pp.25-52.

WARR, P.B. (1966) A serial position effect in the preparation of abstracts. *Language and Speech 9.* 1966, pp. 228-236.