[From: *Meaning: the frontier of informatics. Informatics 9.* Proceedings of a conference jointly sponsored by Aslib, the Aslib Informatics Group, and the Information Retrieval Specialist Group of the British Computer Society, King's College Cambridge, 26-27 March 1987; edited by Kevin P. Jones. (London: Aslib, 1987), p. 151-173.]

SUMMARIZATION: SOME PROBLEMS AND METHODS

John Hutchins
University of East Anglia

Abstract

The provision of summaries is of crucial importance for fully effective retrieval of information, but research on summarization has been relatively neglected, After an outline of the basic linguistic and cognitive complexities of text understanding and summarizing, the paper reviews some current projects towards automating various aspects of summarization, and discusses future prospects.

1. Introduction

Summarization is one of the most common acts of language behaviour. If we are asked to say what happened in a meeting, what someone has told us about another person or about an event, what a television programme was about, or what the latest news is from the Middle East, we are being asked or invited to express in condensed form the basic parts of an earlier spoken or written text or discourse. There are many people whose daily work consists largely, if not exclusively, in the production of condensations or summaries: the journalist who reports on the findings of an inquiry; the judge who sums up the evidence presented in court; the civil servant who provides a survey of arguments for and against a particular proposal; the company secretary who records the minutes of a board meeting; the scientist who reviews recent publications in his research field; and so we could go on.

In linguistics there has been little discussion of summarization. This is a reflection of linguists' neglect, until recent years, of the processes of text production and comprehension. Summarization clearly involves both these still poorly understood processes, and adds a third (condensation, abstraction, generalization). While linguists have paid little attention to this common linguistic activity, this has not been the case for others. The problems of abstracting have long engaged those involved in information retrieval. In recent years cognitive scientists have studied the role of summarization in the recall and remembrance of discourse. Researchers in artificial intelligence have investigated the place of global text organizing features (including summaries) in the understanding of discourse,

In information retrieval the trend in recent years has been towards full text retrieval. In this context it might superficially appear that abstracting has a diminishing role. Abstracts are expensive to produce, requiring both subject expertise and specialist writing skills; since the beginnings of IR automation, researchers have sought for methods of automatic indexing and abstracting. Now, perhaps, it might be thought this is no longer necessary. However, there is plenty of evidence that summaries are still essential for efficient retrieval. The experience of most IR searches is that precision is far higher if free-text searching is performed on abstracts rather than on full texts. This situation is most unlikely to change; authors will continue to express the essence of their messages in summaries, and potential readers will look for these summaries to see whether the text is of likely value to them and worth their while to read.

Summarization involves the understanding of texts. Discourse comprehension, as Van Dijk (Kintsch and Van Dijk 1978, Van Dijk and Kintsch 1983) has demonstrated, necessarily involves the recognition of text structures and the ability to summarize. There is now a considerable measure of agreement about the essential aspects of text structure, although there are many differences in the details of analysis. The following section is an attempt to synthesize the major aspects of text organization from the perspective of summarization.

2. Text structure

A basic distinction is drawn between the microstructure and the macrostructure of texts. The microstructure represents the relations between sequences of sentences in actual text; the macrostructure represents relations between blocks of sentences and the global organization of texts. Relations at both levels may be thematic or semantic. Thematic relations include those of anaphora, backward and forward reference, and topicalization. Semantic relations include those of cause-effect, time, and logical relations.

At the lowest level, text understanding requires the determination of the thematic and semantic progression of sentences and clauses (the microstructure). At a higher level, it requires the construction of global organizational patterns (the macrostructure). Over and above this, full understanding requires the 'integration' of the text message into readers' knowledge bases for the relevant topics (Van Dijk 1979, 1980, Van Dijk and Kintsch 1983).

There is much evidence (Van Dijk and Kintsch 1983) to suggest that what any reader of a text (or listener to a discourse) remembers of its content is something like the semantic network of its macrostructure, i.e. a reader remembers the sequence of the major episodes and the main participants, or in the case of expository text the reader remembers the general gist of the argument and the main points for and against. What is rarely recalled (unless prompted) is the details of microstructure, the particular sequence of events in an episode or the specific stages in an argument, and least of all remembered are the particular surface forms of the text (unless they have a striking literary or metaphorical value).

I will illustrate the basic properties of microstructure and macrostructure and the operations of summarization with an extract from a text used by Van Dijk (1977) and Kintsch and Van Dijk (1978). The text is taken from Heussenstam:

> A series of violent, bloody encounters between police and Black Panther Party members punctuated the early summer days of 1969, Soon after, a group of Black students I teach at California State College, Los Angeles, who were members of the Panther Party, began to complain of continuous harassment by law enforcement officials. Among their many grievances, they complained about receiving so many traffic citations that some were in danger of losing their driving privileges. During one lengthy discussion, we realized that all of them drove automobiles with Panther Party signs glued to their bumpers. This is a report of a study I undertook to assess the seriousness of their charges and to determine whether we were hearing the voice of paranoia or reality.

3. Microstructure

Relations at the microstructural level include anaphora, lexico-semantic links, semantic cohesion and thematic progression.

3.1. Anaphoric and semantic links

Anaphoric links are familiar and need no long explanation (c.f. Halliday & Hasan 1976): the first references to individuals and objects are generally by indefinite expressions (e.g. *a scientist*): subsequent references are generally by definite expressions, e.g. pronouns (*he*), generic nouns (*the man*, *the researcher*), elaborations (*the owner of the white_Volkswagen*), etc. Such links may extend not only across adjacent sentences but also throughout whole texts. Examples of anaphora in the text above are *they*, *their*, *them* referring back to *students*; *we* referring back to *I* and *students I teach*.

Semantic links are represented by the recurrence of lexical items from the same semantic 'field'. Three such fields are clear in the extract (police, complain, transport), represented by the sets:

1) *police, law enforcement officers*
2*) complain, grievance, complained, charges*
3) *traffic, automobiles, bumpers*

3.2. Semantic cohesion

Semantic links between successive sentences or clauses provide local text cohesion. Most familiar are Time relations marked by conjunctions: *before*, *after*. Other relations are Condition-Consequence, Instrument-Achievement, Cause-Consequence, Compatibility, Contrast. These clause relations may be signalled by subordinators (i.e. conjunctions such as *because*, *by...-ing*, *on condition that*, *so that*), sentence or clause conjunctors (*accordingly*, *basically*, *besides*, *in general*, *however*, *therefore*), and by lexical items (e.g. *situation*, *question*, *answer*, *fortunately*, *difference*, *comparison*). Subordinators constitute a finite 'closed' set in languages; the conjunctors a larger but still fairly restricted set; the last group is much more 'open' but nevertheless still represents a relatively homogeneous and identifiable set of vocabulary items (Winter 1977, Hoey 1983).

An example of two clauses linked by items from the three types of markers (from Winter 1977):

> (1): *By* appeal*ing* to scientists and technologists to support his party, Mr Wilson won many middle class voters in the election.
> (2): Mr Wilson appealed to scientists and technologists to support his party. He *thus* won many middle class voters in the election.
> (3): Mr Wilson's appeals to scientists and technologists to support his party were *instrumental* in winning many middle class voters in the election.

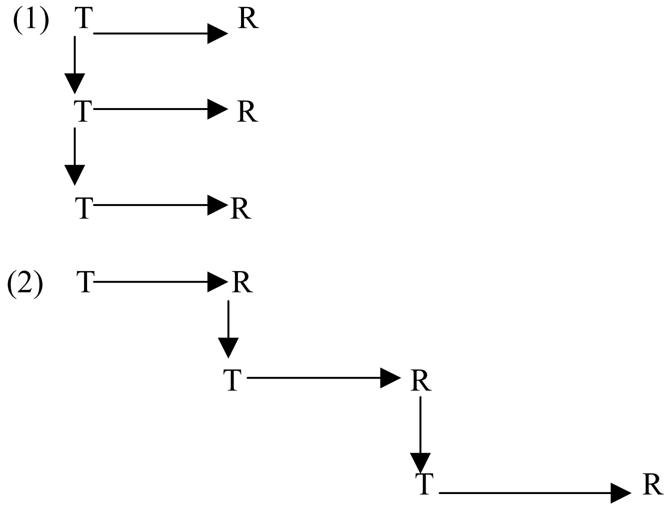
Items of the sets interact closely, e.g.

> They *differed* radically in their approaches to life. He was frugal and deeply religious, *but* she was spendthrift and frankly worldly in all things.

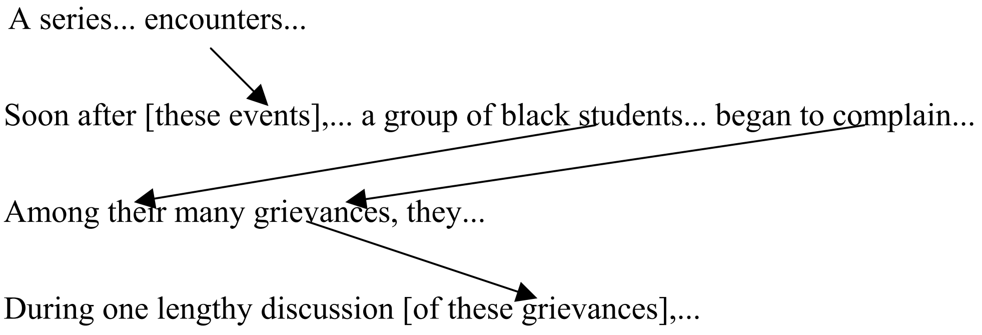
The lexical item *differ* anticipates a contrast, which is contained in the conjunction of two clauses linked by the conjunction but. Such anticipations can extend over longer stretches of text. In this way lexical items (like *difference*, *problem*, *comparison*) can also serve as macrostructural signals indicating relations among paragraphs and whole texts.

3.5. *Thematic progression*

A further aspect of microstructure is thematic progression, the way in which 'new' information is conveyed in the context of what is already known or what has already been expressed. This aspect of text structure is also familiar. In general, the initial noun (phrase), the 'theme' of a sentence (or clause), relates to what has preceded (by an anaphoric link) and the subsequent parts of the sentence, the 'rheme', convey new or unpredictable information. The theme may relate to the theme of the previous sentence (clause) or to an element of the preceding rheme (Daneš 1974, Hutchins 1977). There are thus two basic types of thematic progression: (1) parallel, where the theme is constant, and (2) linear, where themes relate to preceding rhemes:



The thematic progression of the example text:



A series... encounters...

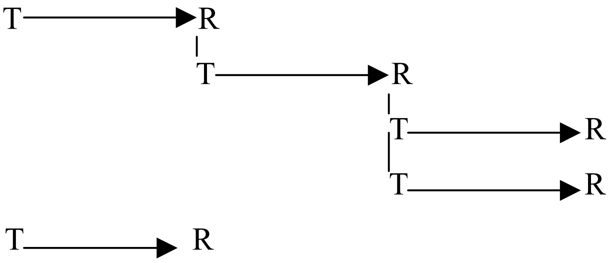
Soon after [these events],... a group of black students... began to complain...

Among their many grievances, they...

During one lengthy discussion [of these grievances],...

This is a report of a study...

Schematically in terms of themes and rhemes:



In this thematic progression there is a break at sentence 5, with two indefinite nouns (a report, *a study*) starting a new progression. It should be noted also that in two cases the thematic links are implicit, from sentence 2 to sentence 1 and from sentence 4 to sentence 3. Ellipses of this nature are common.

Combining the notions of semantic cohesion and thematic progression we may define a 'paragraph' as a coherent text segment focussed on a single topic, comprising a single thematic progression where the theme of the first clause represents the paragraph topic. Such 'microstructural paragraphs' are often similar to conventional written paragraphs but, as the example above shows, they are not identical,

4. Macrostructure

The macrostructure of a text is derived from generalizations of the microstructural propositions (clauses) and deletions of irrelevant or redundant propositions (Kintsch and Van Dijk 1978, Van Dijk 1980). In the selection of 'macropropositions' readers are aided by authors' headings, titles, and summaries, which may be signalled by: *to be brief*, *in short*, *my topic will be...*; and by signals of what the author wants to highlight: *primarily*, *it should be stressed that...*, *I repeat...*, etc.

The macrostructure of a text is organized by a 'macrosyntax' of sequences of episodes or of points of argument, i.e. story plots, scripts, global patterns of problems, solutions, evaluations, etc, Such macrostructural schemata are inferred by readers during their understanding (analysis) of the initial sentences of texts, and are tested and modified continuously against texts as they are read. The recognition of such schemata naturally depends on previous encounters (experiences) with similar texts.

4.1. Narrative macrostructures (story schemata)

The earliest suggested 'macrosyntax' was that proposed by Propp (1968) for Russian folktales. Propp's analysis was in terms of invariant sequences of episodes: the villain's treachery, the hero's rescue of the victim, etc. Later more abstract schemata for narratives were proposed: for example, that of Longacre (1974), who established the sequence Aperture, Setting, Inciting Moment, Developing Conflict, Climax, Denouement, Final Suspense, Closure.

Some of these ideas for 'story grammars' were taken up by researchers in cognitive science and artificial intelligence who saw applications in programs for understanding and summarizing texts. Rumelhart (1975), for example, analysed stories into hierarchical structures. At the top level was a division into Setting and Episode. Setting was divided into a sequence of State descriptions; Episode into an Event and a Reaction to the event. An Event was then analysed as either an episode, a change of state, an action carried out, or a sequence of Events. Reactions were analysed as Internal Responses (e.g. Emotions, Desires), or Overt Responses. Relations between components were to be expressed as relators such as CAUSE, ALLOW, INITIATE, MOTIVATE, etc. For example, INITIATE (Event, Reaction) expressed the relationship between an external event and the main participant's reaction to it. For each relation Rumelhart proposed a 'summarization rule' by which a coherent summary of the story could be generated from the hierarchical analysis.

In practice, AI researchers have not found ways of applying global narrative schemata but have instead sought organizational rules at lower levels, e.g. Lehnert's plot units, and Schank's scripts. These will be considered later.

4.2. Schemata of expository text

It is clear that expository texts have quite different macrostructures. A very common one is the 'hypothesis-testing' or problem-solution type. A typical outline of a scientific paper would be as follows (Hutchins 1976):

Current hypothesis/paradigm
Demonstration of inadequacies
Statement of 'problem'
Statement of new hypothesis or of alternative hypotheses
Testing of hypotheses
Proof of hypothesis or of one of hypotheses
Implications of 'solution'

The two basic components are the statement of a 'problem' and a proposed 'solution'. The scientific paper is 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 researcher is to find a solution or to suggest where one may be found. While the actual processes of research may not conform to neat patterns, the conventions of scientific documentation oblige researchers to present their findings in ways that permit easy assimilation. The author begins therefore by outlining the current approach to the particular matter of interest, reviewing the evidence and the deficiencies of the current view, and concluding with a statement of the precise nature of the problem to be resolved. The core of the paper is the author's proposed 'solution', a description of the tests which have been made and their results, and arguments for accepting the proposed new 'model'. In conclusion, the author may offer comments on the implications of the 'solution' within a wider scientific context.

This schema for expository texts may be interpreted as continuing a long rhetorical tradition (c.f. Kinneavy 1971). It conforms in fact to general patterns of text organization which underlie a very wide variety of text types, from advertisements to technical manuals, and includes narrative texts. There appear (Hoey 1983) to be a limited number of basic patterns, each signalled by discourse markers of the kind recognized by Winter (section 3.2 above).

The most basic pattern is that of a Situation and its Evaluation: if the situation is satisfactory then the evaluation is positive, if the situation is unsatisfactory there is a Problem, a 'negative evaluation of situation'. The Response to the problem is to look for a 'solution' which can be evaluated positively; there may be a number of attempts. Other common patterns are those of a Generalization followed by an Example, a General statement followed by a more Particular statement, a general Preview followed by a discussion of Detail, and a Compatibility pattern in which different situations are compared and contrasted.

As Hoey notes, not all patterns are explicitly marked by discourse signals. The normal interpretation of an unmarked sequence is in terms of chronological plausibility and presuppositions: a Solution presupposes a previous Problem, a Problem is placed in a context (Situation), an Evaluation presupposes a Situation or Solution to be evaluated. Thus, the sequence Situation, Problem, Response (or Solution), Evaluation does not need to be overtly marked, although good style would recommend that it is.

I will illustrate the patterns with the example text. The first sentence presents the general Situation (implicitly, since there is no marker). The existence of a Problem is indicated in sentence 2 by the occurrence of *complain*. For a complaint to be made there must be something wrong in the Situation presented in the first sentence. The problem is described in general terms in the remainder of the sentence as *continuous harassment by law enforcement officers*. Sentence 3 gives a specific Example of the Problem: *Among their many grievances, they complained…..* In sentence 4 there is a preliminary Evaluation of the problem, signalled by the words *discussion* and *realized*. Sentence 5 expresses the general Response to the problem: *a study*, which has as its objective an evaluation (signalled by *assess the seriousness*) in order to decide between alternative explanations (*paranoia or reality*); the latter may be regarded as a Particular aim within the General purpose of the study. Finally, the following text (in a new paragraph) begins to outline the basis for the testing of the hypotheses which is to take place. This is necessarily only an outline sketch of the analysis; its Justification can be derived from Hoey (1983) and Jordan (1984). The resulting schema is:

Situation	(1)	
Problem: General	(2)	
Example of problem	(3)	
Evaluation of problem	(4)	
Response to problem: General	(5a);	Particular (5b))

5. Summarization

Summaries are expressions of the macrostructure of a text as interpreted by an individual in the light of background knowledge. There is consequently no unique summary for a text. In theoretical terms, summarization involves four components: the comprehension of microstructure, the identification of global schemata (microsyntax), the application of macro-rules to generalize and condense to a macrostructural representation (set of macropropositions), and the expression of the macrostructure as a coherent text. As with microstructures, macrostructural representations and the summaries should conform to the basic principles of thematic and semantic cohesion.

5.1. Macro-rules

There are three macro-rules (Van Dijk 1979, 1980), with an additional composite macro-rule:

1) Deletion

Peter saw a blue ball

→ Peter saw a ball

(micropropositions: Peter saw a ball. The ball was blue)

The deletion rule allows for the omission of 'accidental' properties; in general it states that attributes can be omitted in macrostructures on condition that the deleted proposition is not a necessary presupposition of some other following macroproposition; otherwise the coherence of the macrostructure is violated. Deletion may involve complete clauses:

Harry saw a blonde. She was wearing a white frock

→ Harry saw a blonde

2) Generalization

a) *Peter saw a hawk*

→ Peter saw a bird

b) *Peter saw a hawk. Peter saw a vulture*

→ Peter saw birds

c) *Harry watched a tall slim blonde*

→ Harry saw a pretty girl

Whereas the first rule allows for the deletion of accidental properties, the generalization rule allows for the abstraction of 'necessary' properties. In the following examples, hyponyms are substituted (*hawk, vulture → bird; slim. blond → pretty girl*).

3) Construction

Peter bought bricks, sand, cement, laid foundations, erected walls,...

→ Peter built a house

The construction rule is the syntagmatic counterpart of the generalization rule: the resulting macroproposition is entailed by a sequence of micropropositions.

4) Composite deletion-construction:

Peter intended to buy a house... Peter bought a house

→ Peter bought a house

Peter lighted his pipe... Peter smoked a pipe

→ Peter smoked a pipe

The rule may be formalized in terms of actions and their preconditions or in terms of goals and plans. For macrostructures only overall goals and basic actions need to be explicit; it is unnecessary to retain statements of preconditions or particular planning steps.

Van Dijk defines macro-rules in semantic terms, but it may be noted that the deletion macro-rule could also be applied to thematic progressions, i.e. for the omission of propositions which elaborate the rheme of a topic proposition.

Peter saw a hawk. It was hovering above the motorway. Suddenly it swooped on its prey, a large rat.

→ *Peter saw a hawk*

In this sense the elaborations of rhemes are seen as redundant for macrostructures; although it is doubtful that this is invariably true.

5.2. Examples of macro-rule operations

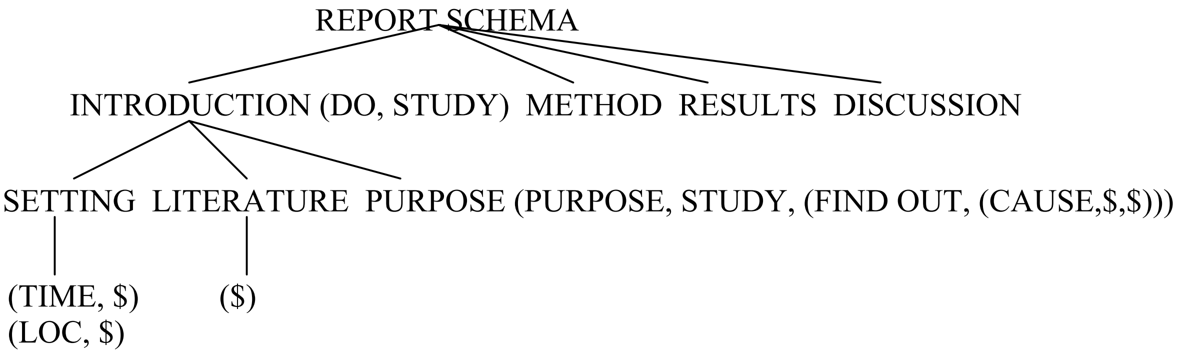
Some of the macro-rules in operation can be illustrated with the example text - as Van Dijk does (Van Dijk 1977; Kintsch and Van Dijk 1978). In sentence 1, unnecessary attributes are deleted: *series of*, *early summer days of 1969*; but some attributes are essential in the context of later text: *Black Panther Party*. In the same sentence, *bloody* is subsumed under the more generic *violent*, and thus omitted. In sentence 2, the attributive phrase *I teach at California State College* is deleted, but *Los Angeles* is retained as a more generic location. In the same sentence, *complained* is omitted since this expression is considered a consequence of *harassment*. And in the following sentence, *in danger of losing their driving privileges* is omitted likewise, as the phrase is regarded as a consequence of *traffic citations*.

The result of these and similar deletions is a 'long summary' comprising the following macropropositions (where referential links are indicated by letters):

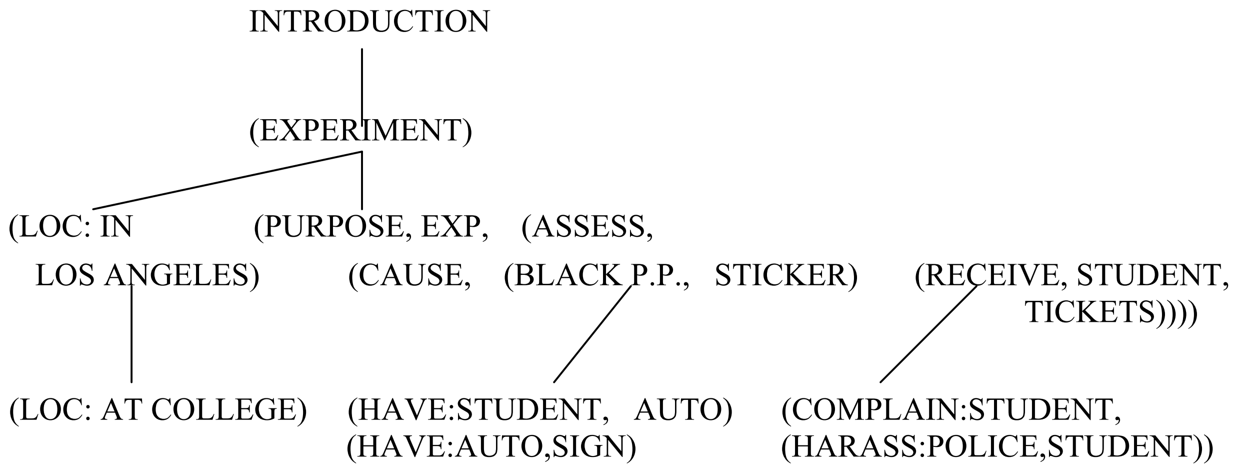
(Ml) There were encounters between police (A) and Black Panther Party members (B)
(M2) Soon after, police (A) harassed students (C) who were Black Panther Party members (B) by (M3)
(M3) Police (A) gave many traffic citations to (C) (M4) Probably (M3) resulted from Black Panther Party bumper stickers on cars of (C)
(M5) We made a study with the purpose of confirming (M4)

This is Van Dijk's analysis (1977); there are a number of uncertainties about precisely how his macro-rules operate, although the general drift of the process is clear. Nevertheless, we may question the validity of some of the operations, in particular the omission of *complained* from sentence 2 (macroproposition 2): even if complaints can be presupposed in cases of harassment, it does not follow that complaints were actually made: the fact that they *were* made is an important feature of this particular text.

Van Dijk stresses that macro-rules are guided by text schemata. These are hypothesized early in text interpretation and confirmed or revised in the light of subsequent text. In this case the reader may infer a Report schema:



It is confirmed by the plausibility of inserting the macropropositions. The resulting macrostructure is therefore:



6. Towards automation of summarization

It is both convenient and traditional to distinguish three types of summaries: indicative, informative, and evaluative. Indicative abstracts have a referential function; they record only the essential topics of a text and no details of results, arguments, and conclusions. Informative abstracts are intended as substitutes for the original, recording relevant quantitative and qualitative details (data, methods, hypotheses). Evaluative summaries go beyond the 'objective' description of topics, methods, and data and assess the import of the original in the context of other work in the field. In this regard, evaluative summaries approach the function of critical reviews; indeed, articles reviewing the state-of-the-art in a research field consist primarily of sets of evaluative summaries. Such are the complexities of summarization, however, that virtually all rules and models have concentrated so far on indicative summaries.

6.1. Automatic extracting

The earliest attempts at automatic abstracting were based on devising rules to extract sentences from texts which would jointly express the main ideas of a document. In nearly all cases they were based on statistical approaches: sentences were identified on the basis of high frequency words (excluding function words and items of common vocabulary), e.g. Luhn's pioneering work (Luhn 1958). The results were neither particularly good condensations nor very coherent texts.

Applying statistical selection to the example text would probably identify sentences containing the words *black*, *car*, *cars*, *citation*, *cited*, *driver(s), driving*, *stickers*, *student*, as a result of which sentence 3 would probably be extracted, but possibly no others (not a satisfactory summary!).

Later auto-extraction systems combined more sophisticated statistical methods and the use of textual 'cues' to identify important (topical) passages, i.e. to act as quasi 'deletion' rules (in Van Dijk's terms), Edmundson (1969) and Rush et al. (1971) used three types of cues:

(1) the recurrence of words in titles and section headings, or the occurrence of words synonymous with them. It is obviously based on the common-sense observation that these are valuable guides for summarization. The example text is entitled 'Bumper stickers and the cops'; application of this cue would tend to highlight sentences with *police* or synonyms and *bumpers*, *automobiles*, *cars*, etc. (assuming generic identification).

(2) the presence of words indicating the importance for authors of the information presented (e.g. *significant*, *primarily*, *impossible*). This cue is clearly related to the kinds of

discourse signals investigated by Winter (1977) and Hoey (1979). A possible candidate in the example is *realized* in sentence 4 (which is not satisfactory on its own as a 'summary' of this paragraph).

(3) the location of sentences within paragraphs and sections. The idea of applying the deletion rule to thematic progressions has been mentioned earlier (section 5.1.). In the example text, its application would isolate as 'topic sentences' the sentence 1 and the first clause of sentence 5, producing a rather unsatisfactory summary by omitting the 'problem' statement.

6.2. Thematic selection

The basic argument for thematic selection as a basis for summarization processes is that the topic sentences of 'paragraphs' express the essential content, since they represent the starting point for the message of the paragraph. The subsequent sentences are elaborations. (Evidently this approach can produce only indicative abstracts, not informative ones.)

Strict application would limit selection to the clause expressing the thematic starting point, which as we have seen gives an unsatisfactory result. A better motivated approach to thematic selection is Hoey's (1983) suggestion that 'a reasonable skeleton summary...can be achieved by the simple expedient of taking the first sentence of each element of the pattern as long as we exclude the signalling clauses'. This means that summaries would be based on the 'topic sentences' (thematic origins) of those sections identified as the Problem and the Response, i.e. the core components of expository texts.

For the example text a summary would be based on sentences 2 (= Problem) and 5a = Response in General) modified to exclude discourse signals and to exclude irrelevant elements. The modifications would Involve the application of Van Dijk-type rules: a deletion rule to eliminate unessential attributes, to convert *began to complain* to *complained*; the use of a synonym rule to replace *law enforcement officers* by *police*; and a rule to delete the discourse signal *soon after* in sentence 2. From sentence 5 would be included only the general statement of the purpose of the study, the Particular hypotheses would be omitted:

> A group of black students who were members of the Black Panther Party complained of harassment by police. This is a report of a study to assess the seriousness of their charges.

The result is a reasonably satisfactory summary.

6.3. Construction rules and scripts

Of Van Dijk's three macro-rules the one explored most thoroughly so far has been that of construction. It is a consequence of considerable AI research on the understanding of narrative texts, in particular by Schank and his colleagues at Yale University.

Most pertinent in this context is the research of Schank & Abelson (1977) on 'scripts' (outline sequences of events or actions to be expected in particular situations), e.g. the series of activities in a restaurant, in a traffic accident, in kidnappings, etc. It was found that problems of interpretation (resolution of semantic ambiguities, identification of anaphoric relations, etc.) are greatly simplified if text passages can be matched to a standard 'script'. Scripts are causal chains representing stereotypic actions in a given setting. In constructing representations of 'messages' such scripts supply predictions and expectations, allow 'common-sense' inferences to be made, and so forth.

Later Schank (1982) proposed components of scripts, Memory Organization Packets (MOPs), representations of knowledge common to many different situations. For example, where a 'script' for visiting a doctor would specify *a* sequence - having a medical problem, making an appointment, going to the surgery, sitting in *a* waiting room, having treatment, etc. - there would be a MOP for all 'professional visits' to lawyers, doctors, dentists, etc., a MOP for travelling, a MOP for making appointments, a MOP for waiting for a service, a MOP for fixing problems, etc. The 'superscript' for visiting a doctor would then be constructed from these and other MOPs.

Further modifications have resulted in the introduction of representations of intentionalities (plans) and in the positing of goal-directed organizations of episodes (Schank 1982): Thematic Organization Points (TOPs), i.e. event sequences which are stored in memory as wholes.

Also relevant is Lehnert's (1982) notion of 'plot units': propositions at the microlevel are instantiations of patterns constructed from atomic elements (Positive events, Negative events, Mental states), linked by four types of relation (Motivation, Actualization, Termination, Equivalence). Thus, for example, a 'plot unit' expressing SUCCESS would be a Mental state actualized by a Positive event; for FAILURE the actualization would be a negative event; a PROBLEM is expressed as the Mental state motivated by a Negative event; the RESOLUTION of such a negative event is through its Termination by a Positive event. The analysis of a story in terms of plot units results in a complex network where some units are subordinated to others. For Lehnert the summarisation of a story involves essentially the extraction of the top-level plot units.

Related to both TOPs and plot units is the notion of Thematic Abstraction Units (TAUs) implemented in Dyer's AI program BORIS (Dyer 1983). TAUs are representations of the moral or point of narratives (often expressed as adages such as 'the pot calling the kettle black'). They incorporate an abstracted intentional structure which represents the pattern of a situation outcome in terms of the plan used, the intended effect, why it failed, how failure could be avoided, etc.

BORIS attempts to integrate different parsing approaches, integrating both microstructural information and macrostructural predictions and expectations. The first integrated parsers developed by the Yale group were the text skimmers. Rather than trying to understand everything, text skimmers sought to identify the gists of stories. The best known example is FRUMP (DeJong 1982) which can recognize over 50 story types and produce brief summaries of the main events.

6.4. The FRUMP summarization program

The experimental program FRUMP (DeJong 1982) works from 'sketchy scripts' of typical newspaper stories (kidnaps, acts of terrorism, diplomatic negotiations, etc.). It skims through texts looking for words signalling a known 'script', for which it is able to predict or expect the occurrence of other words or phrases and so build up the outline of the story. It is only 'interested' in, and only interprets, those parts of texts which relate directly to elements of a 'sketchy script'; the rest of the text is ignored or 'skipped'. (It is not an implausible model of the single-tracked newspaper reader only interested in, say, reports of football games.)

The operation of a script understander can be illustrated by the analysis in Schank et al. (1980) of this passage:

> An Arabic speaking gunman shot his way into the Iraqi Embassy here (Paris) yesterday morning, held hostages through most of the day before surrendering to French policemen and then was shot by Iraqi security officials as he was led away by the French officers.

The first three words are skipped (although stored for later reference if needed). The fourth word *gunman* is identified as a 'high interest actor' which prompts requests for information from the text; (a) who is he? – causing a search for adjectives related to this noun; (b) what did he do? – predicting that he *shot* someone and requiring confirmation; (c) who did he shoot? - creating interest in the verb's syntactic object; (d) why did he shoot? - causing a search for a reason; (e) where did this happen? – causing a search for a location; finally *gunman* prompts searches for the instantiation of one of the 'scripts' ROBBERY, TERRORISM KIDNAP. These questions now guide the process of understanding the story; the next word *shot* confirms the prediction of what the gunman did; *Embassy* provides the location and, as a place of political significance, instantiates the TERRORISM script and sets up further questions about the taking of 'hostages', demands for money, measures to counteract the terrorism, etc. The occurrence of *hostages* confirms the TERRORISM script and allows the potentially ambiguous held (which had been skipped) to be readily interpreted. Parsing continues in this way producing finally an outline (summary) representation:

```
$TERRORISM                          UNEXPECTED RESULT
ACTOR    Arab gunman
PLACE    Iraqi Embassy              $SHOOT
SCENES                              ACTOR    Iraqi  officers
$HOSTAGES    some                   OBJECT    Arab gunman
$CAPTURE                            RESULT
ACTOR    French policemen           STATE    dead
OBJECT    Arab gunman               ACTOR    Arab gunman
PLACE    Iraqi  Embassy
```

The interpretation of shot by Iraqi security officials as an unexpected result arises because the TERRORISM script had already been completed with the surrender to French policemen. The second occurrence of shot in the last sentence prompts the expectation of a new script, which is only partly instantiated by what follows.

The following is an example of a FRUMP summary using similar methods of analysis.

The source text:

> A small earthquake shook several Southern Illinois counties Monday night, the National Earthquake Information Service in Golden, Colo., reported.
>
> Spokesman Don Finley said the quake measured 3.2 on the Richter scale, "probably not enough to do any damage or cause any injuries." The quake occurred about 7:48 p.m. CST and was centered about 30 miles east of Mount Vernon, Finley said. It was felt in Richland, Clay, Jasper, Effington and Marion Counties.
>
> Small earthquakes are common in the area, Finley said.

Summary:

> THERE WAS AN EARTHQUAKE IN ILLINOIS WITH A 3.2 RICHTER SCALE READING

The example illustrates that FRUMP picks up only those details which it is 'interested' in, i.e. which it is programmed to search for. Having identified the EARTHQUAKE script it has an interest in Richter readings but it *is* not interested in exact locations; nor, in particular, whether earthquakes are common or not in the area. (It may be noted that *a* Hoey-type summary (6.2 above) might not miss this Evaluation statement.)

In the case of our example text (section 2), we can envisage that a FRUMP-like approach might involve 'sketchy scripts' on black student protests, harassment, and traffic violation. A HARASSMENT script (or MOP) prompted by the occurrence of *harass* would presumably accept *police* officers as potential agents and students as 'victims', and would predict that *complaints* would be made. A script for TRAFFIC VIOLATION would include as one prediction that offenders might *lose their driving privileges*. And so forth.

As an attempt to automate what is in effect Van Dijk's macro-rule of construction, FRUMP clearly represents an important step forward in understanding summarization. It is also a reasonably plausible model of certain kinds of specialized abstracting where only those parts of documents are analysed and recorded which are of interest to a particular research project. In other words, it suggests summarization programs tailored to specific user needs.

6.5. The SUSY summarization program

At the University of Udine in Italy an experimental summarization system is designed to produce user-specified summaries (Fum et al. 1982). In computer dialogue interaction the user will define a 'working text schema' outlining vocabulary and content requirements, and a 'working summary schema' specifying the type of summary required. A library of possible schemata is envisaged. In processing a text, the parser constructs a representation, essentially of micropropositions of the Van Dijk kind. This representation is then to be converted into a 'weighted' representation using the 'working text schema' - i.e. the 'interesting' parts of the original are selected (as in FRUMP). From this selection the 'working summary schema' is to generate the desired summary. How much of the project has reached practical realization is not known.

6.6. An implementation of macro-rules

A small-scale experiment for creating summaries using the Yale techniques, and again based on Van Dijk's ideas, has been the research of Correira (1980) at the University of Texas. The attempt was made to automate Van Dijk's macro-rules, primarily the construction rule using the Yale 'script' approach. The corpus was Van Dijk's analysis of a *Decameron* story (Van Dijk 1979). As in FRUMP, particular scripts were activated on encountering specific lexical items: *purchase* initiated a $PURCHASE script. The macro-rules were thus applied 'bottom-up', macrostructures were predicted from fragments and story schemata were tested 'top-down'. As in Van Dijk's descriptions the higher level macropropositions were held to represent summaries of the story, which could then be generated by a text production program. The importance of the work is that it shows that there are computational methods readily available which can deal with Van Dijk's macro-rules.

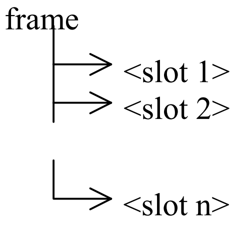
6.7. The TOPIC summarization system

The TOPIC project at the University of Constance in West Germany (Hahn and Reimer 1986) goes further in that it seeks to demonstrate the applicability of AI techniques in a system designed specifically to produce indicative abstracts in the field of information technology. The TOPIC system combines the frame representation approach familiar in many AI projects, a FRUMP-type partial parser, and the organization of text in thematic patterns.

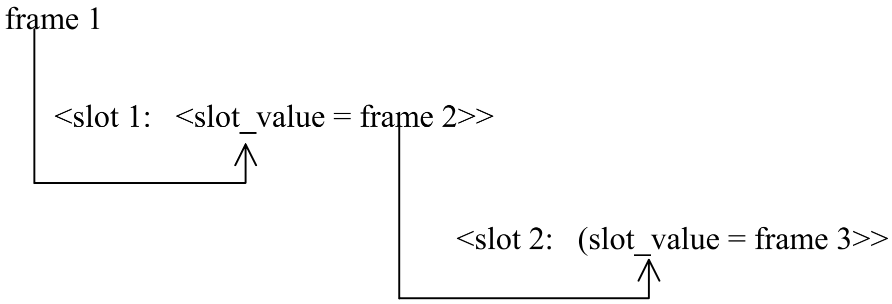
Texts are directly mapped on to frame-slot representation structures, without passing through any intermediate linguistic analyses (e.g. of syntactic structure). The text parser has two main components: a 'world knowledge' database providing the information for interpreting and relating concepts in the field of interest, and a set of decision procedures (word experts) utilizing this knowledge to relate concepts to lexical items and to determine patterns of thematic progression. Frames are stereotyped pre-defined fragments of 'knowledge' (like scripts and MOPs). Word experts (derived from work by Small and Rieger 1982) provide lexical, syntactic and pragmatic information, make predictions and inferences, build interpretations, and take into account lexical cohesion (anaphora) and lexical relationships (thesaural links). As in FRUMP the system picks up only words of 'interest' (in this case names of computers, information on RAM, sizes, bytes, etc.) and seeks to place the information, from the appropriate word experts into relevant frames. Thus it acquires 'new information' from the text by filling slots and instantiating new frames. Analysis is 'bottom-up' from the lexical elements; the system does not refer to or attempt to construct a global schema, it makes no assumptions about text macrostructure.

The result of a parse is a knowledge structure corresponding to the 'message' of a text fragment in the form of a frame representation. At the end of each paragraph the system attempts to condense the frames identified, by determining which frame is to be the 'topic' (based on frequency and connectivity patterns and thematic links indicated by the slot-filling which has occurred), and by subordinating other frames to the 'topic' frame. In the process some redundant frames are eliminated. The result is a frame representation for the whole paragraph. In a final procedure, the paragraph representations are themselves merged ('topic' frames are compared and subordinating links are identified), and a text graph is produced which represents the coherence relations of the text in terms of themes and rhemes. In TOPIC there are three basic frame-slot patterns corresponding to the two types of thematic progression (3.2 above) and a third 'derived theme' pattern: frames are themes, slots are filled by rhematic 'new' information.

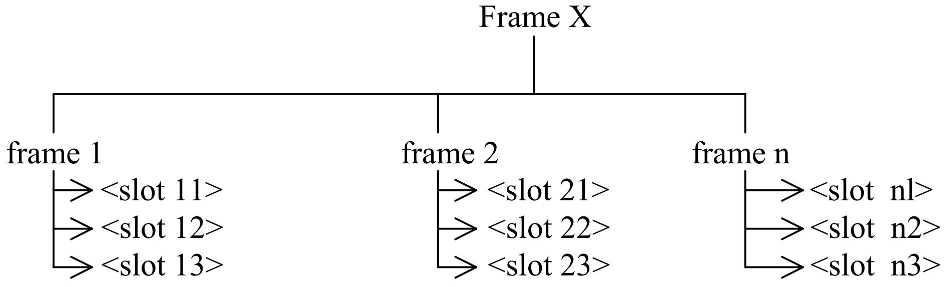
1) Constant theme:



2) Linear thematization:


frame 1
<slot 1: <slot_value = frame 2>>
<slot 2: (slot_value = frame 3>>

3) Derived theme:



The following is an example of the operation of a frame-slot analysis in a paragraph with constant theme:

> The PC2000 is equipped with a 8086 cpu as opposed to the 8088 of the previous model. The standard amount of dynamic RAM is 256K bytes. One of the two RS-232C ports also serves as a higher-speed RS-422 port.

The resulting representation is:

```
PC2000
       <cpu>
       8086
       <main memory>
       RAM
            <size>
            256K bytes
       <port>
       RS-232C
       RS-422
```

The first noun *PC2000* is 'known' to be a microcomputer; its word expert sets up the appropriate frame. Following words in the passage fill slots in this frame, i.e. it is expected that the slots <cpu>, <main memory> and <port> will be filled, and they are with *8086*, *RAM* and *RS-232C* respectively.

Summaries are produced by picking out the highest nodes of a (paragraph or text) frame representation, i.e. those frames identified as themes: *PC2000* only in the above example. As before, the argument is that 'topic' propositions provide good indicative abstracts.

Of all the implementations described TOPIC approaches closest to a genuine summarization program, incorporating knowledge representations, word experts, partial parsing, clustering algorithms, and thematic progression patterns. It is limited, like most experimental systems, to a narrow subject field, and so its feasibility in general is still an open question. However, the approach is promising; it is not devoted to the exploration of a single method (unlike some AI systems), but is prepared to test the feasibility of different aspects. This is essential at the present stages of summarization research.

7. Limitations of present research

From these descriptions it is clear that the two macro-rules of deletion and construction have been most actively investigated. By contrast, the third macro-rule, generalization, has been relatively neglected.

The particularly difficult problems of generalization will not be fully resolved by the use of semantic networks, knowledge databases, word experts and frame representations. Van Dijk's own descriptions (1977, 1979, 1980, etc.) of summarization are imprecise and informal, and rely on intuitive (subjective) judgements of what are plausible substitutions in particular contexts. There is a widespread assumption that generalization must involve reference to an internalized 'thesaurus' (e.g. of the Roget type for general vocabulary) or to an internalized complex of paradigmatic relations (synonyms, polysemes, antonyms, hyponyms, etc.). The TOPIC project is exploring this approach in a narrow subject field. However, while thesaural networks may well be valid for specialized 'hard' vocabularies in, for example, computing and in most sciences, engineering and medicine, they have been found to be both more difficult to establish and less valuable in practice within the 'softer' vocabularies of the social sciences. The difficulties are even more intractable in general (common) vocabulary; but it is above all in this area that generalization has to work well. It seems clear that generalization means more than the consultation of thesaural hierarchies; and much more research is needed on methods of semantic analysis appropriate not only for summarization but for natural language processing generally in the area of information retrieval.

Other problems may be briefly mentioned. Semantic and thematic links within microstructures are often implicit or elliptical; the thematic progression of the example text is a good illustration (3.3 above). Macrostructural organization is poorly understood; AI research has concentrated on narrative texts (Rumelhart 1975, Lehnert 1982, Dyer 1983) and too little work has been carried out on expository texts (DeJong 1982). Research on lexical signals of text organization has been too long neglected, and the work of Winter and Hoey represents just the beginning of what is necessary in this field. Hunston (1985), for example, suggests that despite its considerable plausibility, the basic Problem-Solution pattern needs refinement to deal with actual texts, perhaps as a variety of different Goal-Step patterns, each signalled by distinctive conjunctors *in order to*, *so that*, *so as to*, *contributes to*, *requires*, etc.

Much macrostructural analysis of expository text remains subjective, unformalized, and perhaps at present unformalizable. Any general 'rules' established will always be no more than indicators of possible structural patterns; practical analyses of macrostructures have to be flexible and adjustable to the particular features of specific texts. For summarization, the framework of future studies may well have now been provided by Van Dijk, supplemented by the work on discourse signals and by experimental research in artificial intelligence. However, it will also be essential that the expertise of professional abstractors (and other summarizers) is properly exploited; otherwise there will be a real risk of misguided and misdirected research.

8. Summarization and information retrieval

Writers make assumptions about how much their potential readers know of the subject they are writing about. Often, highly specialized knowledge is presumed, particularly in scientific papers, If a reader lacks any knowledge presupposed by the author (taken as 'given') then he or she will have difficulties. What the writer assumes to be already known is the starting point for the thematic progression of the text, i.e. it is contained in the 'topic' paragraph of the microstructure and the 'topic' propositions of the macrostructure. It is here that the context (Situation) and the questions to be treated (Problem) are expressed; they are the foundations for both the thematic and the semantic progression of the text. Readers who have insufficient knowledge to understand the general situation or to grasp the nature of the problem are unlikely to learn much from the rest of the text.

The aim of IR systems is to put readers into contact with texts with the potential to rectify the gaps and 'anomalies' in their personal states of knowledge (Belkin et al. 1982). In crude terms, IR systems must provide users with documents which start from their particular level of understanding and contribute to their specific 'problem-solving' needs. In an ideal situation, they should be flexible enough to provide tailor-made sets of documents filling the specific knowledge gaps of individual enquirers. In practice, they are designed for general use on the basic of what is assumed to be 'common' knowledge for *a* particular clientele. Summaries intended for a wide potential readership (and most abstracts and review articles are) indicate the major themes (topics) of a text and in so doing indicate the knowledge assumed for its understanding. In this way summaries provide the essential link between semi-structured 'information needs' and full texts of potential relevance; they are recognized as important components in maximizing retrieval effectiveness.

However, summaries are more significant in information organization than only in the form of abstracts. The indexing of documents is a valuable aid to reliable and consistent retrieval, and indexing is itself a variety of summarization (although without the constraints of text coherence). Review and survey articles are recognized tools for evaluating and 'systematizing' current and past research in scientific disciplines. They consist almost exclusively of summaries of primary reports of research: sometimes very brief, sometimes simply citations (surrogates of summaries), often merely indicative but frequently informative or evaluative, Reviews of individual monographs are likewise essentially indicative and/or evaluative summaries. Then there are encyclopaedia articles, handbooks, textbooks and popularizing articles - all varieties of summaries. Finally, the most recent tools of knowledge organization are the 'expert systems' derived from AI research. These consist of summaries of the consensus knowledge and the agreed methods for making decisions, within a relatively narrow subject field. It can indeed be argued that all 'common' knowledge within a scientific discipline is made up of 'summaries' (as Van Dijk stresses (2 above), understanding – and therefore knowledge organization – necessarily involves summarization).

Information access and retrieval depends crucially on all these various types and tools of knowledge organization. There are strong reasons, therefore, for more thorough studies of the processes of summarization. It brings together information scientists, linguists, and researchers in artificial intelligence, and many others. The aim should initially be not so much 'automation' but basic understanding. The problem to be tackled is clear: all readers and writers, speakers and listeners are summarizers (and some are specialists); summarizing is essential to both text understanding and to text production, and it is crucial to information and knowledge organization; but we are ignorant about almost every aspect of the processes involved.

References

Belkin, N.J. et al. (1982) 'ASK for information retrieval' *Journal of Documentation* 38, 1982, 61-71; 145-164.

Correira, A. (1980): 'Computing story trees' *American Journal of Computational Linguistics* 8(3/4), 1980, 135-149.

Daneš, F. (1974): 'Functional sentence perspective and the organization of text' in Daneš, F. (ed.) *Papers in functional sentence perspective* (The Hague: Mouton, 1974), p.106-128.

DeJong, G. (1982): 'An overview of the FRUMP system' in Lehnert, W.G. and Ringle, M.H. (eds.) *Strategies for natural language processing* (Hillsdale: Erlbaum, 1982), p.149-176

Dyer, M.G. (1983): *In-depth understanding: a computer model of Integrated processing for narrative comprehension.* (Cambridge, Mass.: MIT Press, 1983).

Edmundson, H.P. (1969): 'New methods in automatic extracting' *Journal of the ACM* 16, 264-285.

Fum, D. et al. (1982): 'Forward and backward reasoning in automatic abstracting' in: Horecky, J. (ed.) *COLING 82* (Amsterdam: North-Holland, 1982), p. 83-88.

Hahn, U. and Reimer, U. (1986): *TOPIC essentials.* Konstanz: Univ., 1986 (Bericht TOPIC-19/86).

Halliday, M.A.K. and Hasan, R. (1976): *Cohesion in English.* London: Longman, 1976.

Heussenstam, F.K. (1971): 'Bumperstickers and the cops' *Transactions* 8, 1971, 32-33.

Hoey, M. (1983): *On the surface of discourse.* London: Allen & Unwin, 1983.

Hunston, S. (1985): 'Text in world and world in text: goals and models of scientific writing' Nottingham Linguistic Circular 14, 1985, p.25-40.

Hutchins, J. (1976): 'On the structure of scientific texts' *UEA Papers in Linguistics 5,* 18-39.

Hutchins, J. (1977): 'On the problem of "aboutness" in document analysis' *Journal of Informatics* 1(1), 1977, 17-35.

Jordan, M.P. (1984): *Rhetoric of everyday English texts.* London: Allen & Unwin, 1984.

Kinneavy, J.L. (1971): *A theory of discourse: the aims of discourse.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Kintsch, W. and Van Dijk, T.A. (1978): 'Toward a model of text comprehension and production' *Psychological* Review 85(1), 1978, 363-394.

Kuhn, T.S. (1970): *The structure of scientific revolutions.* 2nd ed. Chicago, Ill.: Chicago Univ. P., 1970.

Lehnert, V.G. (1982): 'Plot units: a narrative summarization strategy' in Lehnert, V.G. and Ringle, M.H. (eds.) *Strategies for natural language processing* (Hillsdale: Erlbaum, 1982). p.375-412.

Longacre, R.E. (1974): 'Narrative versus other discourse genre' in Brend, R.M, (ed.) *Advances in tagmemics* (Amsterdam: North-Holland, 1974), p.357-376.

Luhn, H.P. (1958): 'The automatic creation of literature abstracts' *IBM Journal of Research and Development 2,* 159-165.

Propp, V. (1968): *Morphology of the folktale.* 2nd ed. Austin, Tex.: Univ. of Texas P., 1968.

Rumelhart, D.E. (1975): 'Notes on a schema for stories' in: Bobrow, D.G. and Collins, A. (eds.) *Representation and understanding* (New York: Academic Press, 1975), 211-236,

Rush, J.E. et al. (1971): 'Automatic abstracting and indexing, II: Production of indicative abstracts by application of contextual inference and syntactic coherence criteria' *Journal of the American Society for Information Science 22,* 260-274,

Schank, R.C. and Abelson, R. (1977): *Scripts, plans, goals, and understanding.* Hillsdale, N.J.: Erlbaum, 1977.

Schank, R.C. et al. (1980): 'An integrated understander' *American Journal of Computational Linguistics* 6, 13-30.

Schank, R.C. (1982): 'Reminding and memory organization: an introduction to MOPs' in Lehnert, W.G. and Ringle, M.H. (eds.) *Strategies for natural language processing* (Hillsdale: Erlbaum, 1982), p.455-493.

Small, S, and Rieger, C. (1982): 'Parsing and comprehending with word experts (a theory and its realization)'. In Lehnert, W.G. and Ringle, M.H. (eds.) *Strategies for natural language processing* (Hillsdale: Erlbaum, 1982), p.89-147.

Van Dijk, T.A, (1977): 'Complex semantic information processing' in: Walker, D.E. et al. (eds.) *Natural language in information science* (Stockholm: Skriptor, 1977), p.127-163.

Van Dijk, T.A. (1979): 'Recalling and summarizing complex discourse' in: Burghardt, W. and Holker, K. (eds.) *Text processing* (Berlin: de Gruyter, 1979), p.49-118.

Van Dijk, T.A. (1980): Macrostructures. Hillsdale, NJ: Erlbaum, 1980.

Van Dijk, T.A. and Kintsch, W. (1983): *Strategies of discourse comprehension.* New York: Academic Press, 1983.

Winter, E. (1977): 'A clause-relational approach to English texts' Instructional Science 6(1), 1977, 1-92.