[From: *UEA Papers in Linguistics* 19, June 1983, p.38-54]

Some Problems and Methods of Text Condensation

John Hutchins
(University of East Anglia)

Introduction

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 to express in condensed form the basic parts of an earlier spoken or written text. We are encouraged to do it from our childhood ("What did the teacher tell you?"); we are required to do it, as students, in examinations ("Outline Plato's arguments in *Phaedo*"); and we are paid to do it in many occupations of later life. There are many people whose daily work consists largely, if not exclusively, in the production of condensations or summaries: the journalist who reports the findings of an inquiry; the judge or defence lawyer 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.

It is surprising that linguists have paid so little attention to such a common linguistic activity (one in which they themselves engage frequently). It is primarily a reflection of their neglect until recent years of the processes of text production and comprehension; summarisation clearly involves both these still poorly understood processes, and adds a third (condensation, abstraction, generalisation). Given its complexity, perhaps the paucity of linguistic research in this area is not so surprising. Some beginnings have been made in text linguistics, in artificial intelligence and in information science (as we shall see), and there are the raw materials for studies of summarisation in codes of practical expertise (viz. manuals and guides for journalists, abstractors, lawyers, etc.).

No attempt can be made in this article to consider all the multifarious forms of text condensation; instead it will be restricted to one area: the production of representations or indicators of the contents of published documents in forms which enable people to discover what documents are 'about' before reading them. It will additionally concentrate on efforts to automate the processes since the application of computerised methods highlights the particular linguistic problems involved in this type of text condensation.

Representations of document contents are found in essentially two different forms, either as coherent text themselves (i.e. abstracts or summaries) or as individual words or phrases not constituting continuous, coherent sentences or texts (i.e. index terms). Abstracts are generally concerned with the representation of whole texts, usually articles in journals. Index terms may refer to parts of texts, to texts considered as wholes, or even collections of texts; hence, the index at the back of a book contains index terms which each refer to short passages in the book, sometimes no more than a single sentence; but the index in a library or a bibliography contains index terms which refer to books or articles as entities. Such differences in indexing specificity reflect the objectives of the indexing services and the needs of the expected users of the indexes. For example, in a specialised information service only those parts of documents would be indexed which are of direct interest to the research or business activities of a small group of users; in a general library, by contrast, indexing would

aim to be disinterested and unbiased, and would treat documents as wholes. Similar factors apply to abstracts; they may either emphasise some parts of a document and ignore others, or attempt to summarise the document as a whole. In addition, it is common to distinguish between 'informative' abstracts which include actual results, figures and conclusions from source documents, and 'indicative' abstracts which simply record the fact that certain topics are covered. Likewise, we may distinguish between 'topic indexing' where the index terms assigned to a document indicate what it is 'about' as a whole, and 'summary indexing' where the index terms record most of the topics covered in the document.

Text structure

This outline of some of the complex factors involved should make it clear that indexing and abstracting are distinct and particular forms of text condensation which cannot be easily subsumed under the general rubric of 'summarisation'. Nevertheless, it may be convenient to start from attempts to identify the general processes of summarisation. In this respect, the work of Van Dijk is of central importance. Within his theory of text linguistics, Van Dijk (1977, 1980) distinguishes between the microstructure of a text (the underlying propositional content of its sentences and clauses, and their connections to each other, in the linear sequence in which they are expressed) and its macrostructure (the semantic representation of the text as an entity, independently of its particular propositional manifestation). Summaries are one way of expressing the macrostructures of texts. Van Dijk suggests that macrostructures are derived from microstructures by the operations of four types of 'macro-rules'. Two are concerned essentially with the identification of 'important' propositions; *deletion* operates negatively by eliminating the unnecessary and irrelevant (e.g. detailed descriptions, background information, common knowledge), and *selection* operates positively by extracting the necessary and relevant (e.g. propositions expressing pre-conditions and data essential for the interpretation of other propositions). The other two are concerned with condensation and abstraction: *generalisation* constructs general propositions from the semantic detail of microstructural propositions (e.g. from a description of girls playing with dolls, boys playing with train sets, etc. it derives a description of 'children playing with toys'), and *construction* replaces sequences of propositions by single propositions expressing self-contained events or processes (e.g. from an account detailing stages in a long journey it derives a simple statement that a journey takes place).

It is widely accepted that macrostructures are themselves organised formally according to general patterns selected to suit the author's (speaker's) purposes. Most familiar are the patterns and principles underlying narrative texts (e.g. Propp (1968), Greimas (1966), Hendricks (1972)), but similar global patterns seem to determine the structuring of expository texts of the kind exemplified by the scientific paper or the scholarly article. One common pattern is the Situation-Problem-Solution-Evaluation type (Hutchins 1977, Hoey 1979, Jordan 1980) in which the author first states the 'current' position on an issue (the 'situation') and points out its inadequacies or defects (the 'problem'), then proposes a 'new' hypothesis or suggests a number of alternative explanations and describes various 'tests' of the new proposals (the 'solution'), and ends by arguing the merits or implications of his proposed 'solution' (the 'evaluation'). These global structures are apparent at the microstructural level (i.e. in the actual texts) in the form of 'discourse signals' which provide readers with cues (or clues) to what they may expect to follow. In narratives, such signals are generally indicators of time relations; e.g. *in the beginning*, *one day*, *later*, *then*,

meanwhile. In expository texts the signals often express 'logical' relations; e.g. *because*, *consequently*, *as a result*, *by contrast*, *in order to*. As Winter (1977) and Hoey (1979) have demonstrated, discourse signals are not restricted to conjunctional and adverbial forms; nouns and verbs such as *achieve*, *addition*, *action*, *attribute*, *basis*, *change*, *compare*, etc. are frequently the bearers of information on the overall structuring of paragraphs and texts.

It is perhaps reasonable to suggest that summarisation represents an attempt to establish (or re-construct) the macrostructure on the basis of discourse signals present in the text (microstructure) and on the basis of the reader's understanding of types of global patterns (text typologies). Other factors are, of course, involved in the process of understanding: general knowledge of the language (in particular the lexis), specific knowledge of the subject of the text and of the lexical usage in that subject, familiarity with other texts of similar natures, etc. The process of summarisation itself may, however, be isolated from such content- (or subject-) dependent particularities as long as we are concerned only with discussion of a theoretical model of summarisation. (The obvious analogy is the attempt by transformational-generative grammarians to devise a theory of syntax independent of the semantic content of sentences and texts.) In practice, summarisation cannot be divorced from an understanding of the content or 'message' of texts; it is obvious that Van Dijk's macro-rules require semantic knowledge, and it would seem unlikely that discourse signals alone are sufficient for text re-structuring.

To illustrate the interdependence of content and structure in text analysis we may take the work on 'information formatting' by Sager and her colleagues (Sager 1978). The goal of the formatting process is to break down a natural language text into standardised categories of information suitable for display in tabular form and for subsequent computer-based statistical analyses. For example, the following medical record would be 'formatted' as in fig. 1:

> Patient first had sickle cell anemia diagnosed at age 2 when he complained of leg pain. He was worked up and diagnosis was made. He was symptomatic until age 5 when he was admitted to Bellevue Hospital with chest pains. He was hospitalized for a month and released.

In the first stage a generalised parser (the linguistic string parser, derived from Zellig Harris' work in mathematical linguistics) produces a phrase-structure, dependency representation of the unedited texts. The second stage involves the segmentation of the parsed output into semantic categories established for the subject areas of the texts in question. In any discipline there are semantic constraints on the acceptability of statements – in effect, the discipline has a 'sublanguage' and its own 'sublanguage grammar'. In cell biology, for example, the statement *the ion crosses the membrane* would be an acceptable proposition (whether true or false in a particular instance), whereas *the membrane crosses the ion* would be rejected as nonsense. Such observations lead to the establishment of subject-specific classifications of vocabulary (e.g. noun-classes such as 'cations', 'enzymes', 'cells', 'proteins') and subject-specific syntactic rules (e.g. $N_{ion}V_{move}N_{cell}$). The notion of 'sublanguages'' is present (sometimes only implicitly) in most artificial intelligence research. In the highly complex information retrieval system devised for the U.S. Navy (Hendrix, et al. 1978), for example, we find subject-specific noun-classes such as 'ship-name' and 'ship-attribute'; and in the well-known natural language understander of Schank and his colleagues (e.g. Schank 1975) we find subject-specific categories embodied in the 'scripts' used for text analysis. The investigation of 'sublanguages' is now

established as an area of computational linguistics (Kittredge & Lehrberger 1982) which may well have considerable importance for linguistic studies in general.

Figure 1

CONJ	PATIENT	TREATMENT		PATIENT STATE					TIME		
		INST	V-MD	V-PAT	BODY PART	NORM	SIGN/ SYMPT	DIAGNOSIS	P1	P2	REF, PT
	patient		first had diagnosed					sickle cell anemia		at	age 2 yrs
when	he			complained of	leg		pain				(age 2 yrs)
	he		was worked up								
and			diagnosis was made								
	he			was		asympto-matic				until	age 5 yrs
when	he	Bellevue Hospital	was admitted to	with	chest		pains				(age 5 yrs)
	he		was hosp-italized						for a month		
and	(he)		(was) released								

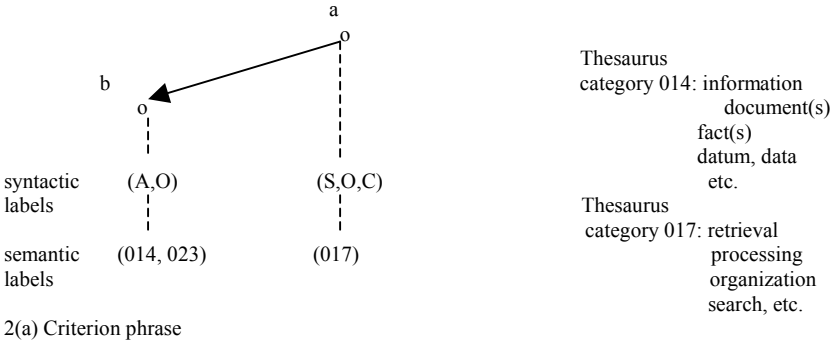
Indexing

Research on automatic indexing has been dominated by the view that summarisation is essentially a matter of vocabulary analysis; most experimental work has concentrated on statistical methods of analysis (Sparck Jones 1974, Harter 1978), the general assumption being that what should be isolated are the 'important' parts of a text and that, broadly speaking, words or phrases which occur frequently are likely to be important indicators of content. However, the crude counting of word tokens is obviously unsatisfactory, and so we find that statistical analyses of texts incorporate many subtle and complex refinements: exclusion of words frequent in a particular subject field (as well as words frequent in general vocabulary, such as function words), truncation of word endings to bring together morphologically related lexical items, normalisation of frequencies to allow for varying text lengths, use of co-occurrence frequencies, and so forth. The methods have proved remarkably successful in practice – people are able to locate the documents they need, without being deluged by masses of irrelevant documents – but there is still the suspicion that the use of some linguistic analysis might lead to improvements. There are two main areas: the analysis of homonyms and of synonyms, and the establishment of anaphoric (in particular, pronominal) relationships.

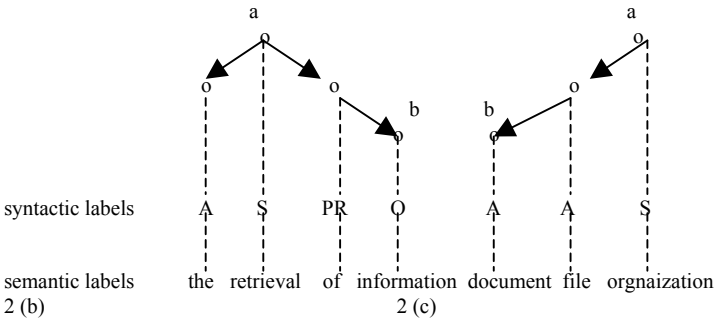
Two projects have attempted some linguistic analysis in automatic indexing: the SMART system and the SYNTOL system. SMART now uses statistical methods almost exclusively, but earlier versions (Salton 1968) used some semantic and syntactic analysis. Each sentence of a document (or, more often, its abstract) was parsed by the Harvard Predictive Analyser (a finite-state parser designed originally in the 1950s for machine translation), which produced a basic phrase structure analysis in dependency grammar format. From the parsing were extracted substructures (e.g. subject- verb and noun-adjective conjuncts) to be matched against a dictionary of 'criterion phrases'. For example, the criterion phrase in fig. 2 (a), representing a cluster of paraphrases of 'information retrieval', would be identified in the phrase structures of figs. 2(b)-(e). The approach encountered considerable problems in the establishment of the syntactic and semantic conditions for criterion phrases, but there

were even more difficulties with the inadequacies of the parser, which notoriously produced either no analyses at all or far too many. It is probable that the more reliable and successful parsers which have been developed subsequently (cf. Grishman 1976) would now produce more satisfactory results; but the 'failure' of SMART has discouraged later research on these lines.

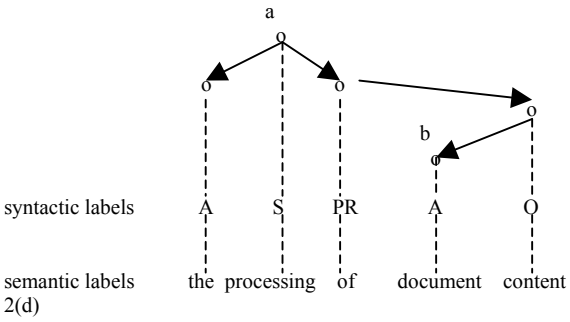
Figure 2



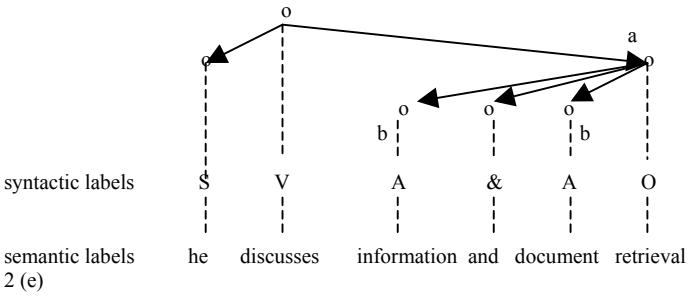
2(a) Criterion phrase



2 (b) 2 (c)



2(d)



2 (e)

Experimentation with linguistic analysis in automatic indexing has also not been helped by the disappointing results of SYNTOL, a system based on a clearly articulated logico-linguistic theory and intended to serve as a general model for linguistics-oriented information retrieval systems (Bely et al. 1970). In certain respects, its analytical procedures were similar to those of SMART: here too, the parser

(context-free) produced from natural language abstracts sets of dependency structures from which substructures could be extracted representing pairs of terms linked by basic (semantico-logical) relations. These were the 'syntagms' which were combined in networks in order to represent the contents of texts. It was realised that in the practical context of an information retrieval system the relationships in syntagms could not be too specific; in fact they were perhaps made too abstract (eventually only three types of links were permitted) and the analysis program was not powerful enough to convert the semantic complexities of the natural language input into the required abstractness of SYNTOL's representations.

In terms of Van Dijk's macro-rules of summarisation the primary operations of indexing are those of 'selection' and 'generalisation'. The statistical methods of automatic indexing concentrate almost exclusively on selection techniques (generalisation being limited to morphological truncation); the selection of terms is typical of the approach referred to above as 'summary indexing', where the indexer attempts to record the 'significant' parts of texts. In SMART and SYNTOL we have seen attempts to automate some semantic generalisation, via 'criterion phrases' and 'syntagms', but it is undoubtedly in manual indexing that generalisation is clearly the principle operation, and it is most obvious in 'topic indexing' where the indexer endeavours to express the global 'aboutness' of a text. It has been suggested that part of this generalisation process is based on clues provided by the theme-rheme articulation of text and paragraph structures and by discourse signals at both micro- and macro-structural levels (Hutchins 1978). The underlying thesis is that writers adopt as their starting point some element or region of knowledge which they may assume their potential readers share, i.e. something is taken as already 'known' or 'given', and this is by and large the 'topic' which is to be discussed or elaborated. Thus, the basic task of the indexer is to identify this topic, around which the whole structure of the text coheres.

Abstracting

If we look at the guides written for abstractors it becomes soon apparent that abstracting is far more than just 'summarising'. In addition to the distinction between 'informative' and 'indicative' abstracting mentioned earlier we find a number of specific recommendations and injunctions; to retain the balance and emphases of the original, to pass no comments (either favourable or critical), to state clearly the purpose of the work described, the methods used and the conclusions reached, and to produce a self-contained coherent text (ideally within a single paragraph) which might stand as a substitute for the original for some purposes.

The need to produce coherent texts has meant that research on automatic abstracting has been more ambitious from the linguistic point of view than automatic indexing. Nevertheless in practice the primary approach has also been statistical. Initially, attempts were made to produce 'abstracts' by the extraction of sentences en bloc from texts on the basis of high frequency words (excluding function words and items of common vocabulary), e.g. Luhn's (1958) pioneering work. The results were neither particularly good condensations nor very coherent texts. Later systems have combined more sophisticated statistical methods (similar to those in automatic indexing) with the use of textual 'cues' to identify important (topical) passages. Edmundson (1969) and Rush et al. (1971) have used three types of 'cues': (i) the recurrence of words in titles, subtitles and section headings, or the occurrence of words synonymous with them, (ii) the presence of such words as *significant*, *impossible*, *hardly*, which indicate authors' views of the importance of the

information presented, and (iii) the location of sentences within paragraphs and sections. The first type is obviously based on the observation that titles, subtitles and section headings tend to express the author's notion of what the topic ('theme') is in the following text. The second type is clearly related to the kinds of discourse signals investigated by Hoey (1979). The third type formalises observations on the occurrence of 'topic sentences' in paragraphs, which are familiar from writings on rhetoric and composition (cf. Christensen 1967) and which Daneš (1974) formalised in terms of theme-rheme articulations at the supra-sentential level of text structure.

Whether based on vocabulary frequencies or textual cues it is clear that these procedures are all instances of Van Dijk's macro-rule of selection. Indeed researchers recognise the limitations of their systems and concede that what they are doing should be more correctly called 'automatic extracting'. However, some procedures have been developed to produce coherent sequences of sentences from those extracted, and these include a certain measure of generalisation. For example, among various refinements by Mathis et al. (1973) of the automatic abstracting system of Rush et al. (1971) we find the alteration of specific references (e.g. *Table 2*, *figure 3*, *the second mechanism*) to general references (*a table*, *a figure*, *a mechanism*) and the conflation of sentences by coordination and subordination. For example, the two extracted sentences:

> The system exceeded the capacity of its present auxiliary equipment. The system was modified for further testing.

could be combined as:

> The system exceeded the capacity of its present auxiliary equipment and was modified for further testing.

This operation involved limited parsing to identify noun, verb, and preposition phrases, to locate antecedents of pronouns, and to recognise parallel structures.

Such simple syntactic manipulations (though not necessarily simple computationally) scarcely touch the real problem of generalisation in summarisation, namely to account for our ability to select a general term or expression to cover the content of a number of more specific expressions. The necessary lexical organisation may well be reflected or modelled in the thesauri (of scientific, medical, technical, etc, terms) which indexers use to guide them in the selection of index terms. In such a thesaurus, for example, *syntax*, *semantics* and *phonology* would be recorded as 'narrower terms' of (more specific than) *linguistics*; *parsing*, *phrase structure* and *complementation* as 'narrower terms' of *syntax*; *semiotics* as a term 'related to' (conceptually) *semantics*; and so forth. In effect, thesauri provide practical examples of the relationships linguists discuss theoretically as hyponymy, hypernymy and other paradigmatic sense-relations (cf. Lyons 1977). Nevertheless, generalisation would appear to be more complex than the selection of lexemes from appropriate levels of a semantic network or hierarchy – a notion which is itself difficult to conceptualise (let alone automate) in the absence of relevant linguistic research. It would seem intuitively obvious that generalisation must also involve the identification of common sense elements in sets of lexemes related syntactically, semantically and pragmatically in relevant ways within coherent text passages. It is true to say that almost nothing is known about these complex operations of information processing, little more than the programmatic speculations of Van Dijk (1977, 1980) and other text grammarians.

Discussion of the remaining summarisation operation, that of construction, can be slightly more concrete in so far as the work within artificial intelligence on story understanding is relevant. Most pertinent in this context is the research of Schank and

his colleagues which was developed from the well known programs involving the use of 'scripts' (outline sequences of events or actions to be expected in particular situations: a familiar example is the 'restaurant script' which sketches the normal action-sequence of calling a waiter, ordering a meal, being served, eating the food, getting the bill, and paying the waiter). 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'; it is, of course, hypothesised that our understanding of all messages (whether linguistic or not) is influenced, and sometimes determined, by expectations of what is normal – by, in effect, our past experience of 'similar' situations. In Schank's story understanding programs, texts are analysed by a conceptual-dependency grammar, which produces a highly complex network representation of text, from which one output is a summary of the story extracted by isolation of the principal sequence of actions (i.e. a kind of 'macrostructural' output). However, it is now proposed by Schank et al. (1980) that equivalent summaries can be produced by text parsers which do not attempt to understand everything in a text and do not need a complete semantic representation. They should therefore be able to deal with new texts for which they have not been prepared, with new vocabulary items, new domains of discourse, new syntactic constructions. An experimental program FRUMP (DeJong 1979) 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', from 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 the text 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, say, in the football results.) Schank et al. (1980) give an example analysis and summary 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: who is he? – causing a search for adjectives related to this noun; what did he do? – predicting that he *shot* someone and requiring confirmation; who did he shoot? – creating interest in the verb's syntactic object; why did he shoot? – causing a search for a reason; 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. 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, as follows:

$ TERROISM	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 satisfied by the surrender to French policemen. The occurrence of *shot* prompts the expectation of a new script, which is only partly instantiated by what follows.

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 the processes of summarisation. There are obvious limitations in the range of stories it can handle – and it would be of great interest to know whether this approach is appropriate for expository texts (as opposed to the simple narratives on which the research has concentrated, as so often in artificial intelligence, e.g. Rumelhart 1975, Lehnert 1982). But it is a reasonably plausible model of certain kinds of specialised 'abstracting' where only those parts of documents are analysed and recorded which are of interest to a particular narrow range of specialised research.

Conclusion

In this brief survey of some methods of summarisation, it is clear that research efforts have nearly always concentrated on single aspects of the complex processes – in Van Dijk's terms, on just one part of one kind of macro-rule. What would appear to be desirable in future research on text condensation is some integration of various lines mentioned in this paper: discourse signalling, statistical methods, thesaural 'generalisation', use of sketchy scripts, theme-rheme structuring, sublanguage grammars, and global text structuring. The route to such an integrated approach (whatever its theoretical merits) will undoubtedly be difficult; it is to be hoped that more linguists recognise the intellectual problems and fascinations of understanding what happens in summarisation. If this entails the abandonment of 'pure' linguistic theory in favour of practical interdisciplinary approaches in linguistic methodology then linguistic studies can only gain by being seen to offer genuine insights in the understanding of linguistic phenomena and not simply abstract formulations of restricted aspects of idealised language.

References

Bely, N., Borillo, A., Virbel, J. and Siot-Decauville, N. (1970). *Procédures d'analyse sémantique appliquée à la documentation scientifique*. Paris: Gauthier.

Christensen, F. (1967) *Notes towards a new rhetoric*. New York: Harper & Row.

Daneš, F. (1974) 'Functional sentence perspective and the organisation of text'. In: F. Daneš (ed.). *Papers in Functional Sentence Perspective* (The Hague: Mouton)' 106-118

DeJong G. (1979) 'Prediction and substantiation: two processes that comprise understanding'. In: *IJCAI-79: Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, Tokyo 1979. (Stanford, Ca.: Stanford Univ.), 217-222

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

Greimas, A.J. (1966) *Sémantique structurale: recherche de méthode*. Paris: Larousse.

Grishman, R. (1976) 'A survey of syntactic analysis procedures for natural language' *American Journal of Computational Linguistics*, microfiche 47

Harter, S.P. (1978) 'Statistical approaches to automatic indexing' *Drexel Library Quarterly* 14, 57-74

Hendricks, W.O. (1972) 'The structural study of narration: sample analyses' *Poetics* 3, 100-123

Hendrix, G.G., Sacerdoti, E.D., Sagalowicz, D. and Slocum, J. (1978) 'Developing a natural language interface to complex data' *ACM Transactions on Database Systems* 3, 105-147

Hoey, M. (1979) *Signalling in discourse*. (Discourse Analysis Monograph, no. 6). Birmingham: Univ. Birmingham, English Language Research.

Hutchins, W.J. (1977) 'On the structure of scientific texts' *UEA Papers in Linguistics* 5, 18-39

Hutchins, W.J. (1978) 'The concept of 'aboutness' in subject indexing' *Aslib Proceedings* 30, 172-181

Jordan, M.P. (1980) 'Short texts to explain problem-solution structures – and vice versa' *Instructional Science* 9, 221-252

Kittredge, R. and Lehrberger, J. (1982), eds. *Sublanguage: studies of language in restricted semantic domains*. Berlin: de Gruyter.

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

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

Lyons, J. (1977) *Semantics, vol. 1* Cambridge: Cambridge Univ.P.

Mathis, B.A., Rush, J.E. and Young, C.E. (1973) 'Improvement of automatic abstracts by the use of structural analysis' *Journal of the American Society for Information Science* 24, 101-109

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

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

Rusch, J.E., Salvador, R. and Zamora, A. (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

Sager, N. (1978) 'Natural language information formatting: the automatic conversion of texts to a structured data base' *Advances in Computers* 17, 89-162

Salton, G. (1968) *Automatic information organization and retrieval*. New York: McGraw-Hill.

Schank, R.C. (1975) *Conceptual information processing*. Amsterdam: North-Holland.

Schank, R.C., Lebowitz, M. and Birnbaum, L. (1980) 'An integrated understander' *American Journal of Computational Linguistics* 6, 13-30

Sparck Jones, K. (1974) 'Automatic indexing' *Journal of Documentation* 30, 393-432.

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

Van Dijk, T.A. (1980) *Macrostructures: an interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, N.J.: Erlbaum.

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