A COGNITIVE PROCESS MODEL OF DOCUMENT INDEXING

JOHN F. FARROW

Department of Library and Information Studies
Manchester Polytechnic, Manchester M15 6BH

Classification, indexing and abstracting can all be regarded as summarisations of the content of a document. A model of text comprehension by indexers (including classifiers and abstractors) is presented, based on task descriptions which indicate that the comprehension of text for indexing differs from normal fluent reading in respect of: operational time constraints, which lead to text being scanned rapidly for perceptual cues to aid gist comprehension; comprehension being task oriented rather than learning oriented, and being followed immediately by the production of an abstract, index, or classification; and the automaticity of processing of text by experienced indexers working within a restricted range of text types. The evidence for the interplay of perceptual and conceptual processing of text under conditions of rapid scanning is reviewed. The allocation of mental resources to text processing is discussed, and a cognitive process model of abstracting, indexing and classification is described.

1. INTRODUCTION

IT HAS OFTEN BEEN OBSERVED that published accounts of indexing either ignore or at best skate gingerly over the mental processes that take place between the indexer's scanning of a document and the writing down of a set of indexing terms. Jones [1] for example has remarked that 'the activity of indexing, as typified by Collison [2], Knight [3] and to an extent by Borko and Bernier [4] tends to be concerned with the mechanics of alphabetisation, cross-referencing and the form of . . . index entries. The relationship between text and index entries is rarely examined'. There has, since Jones wrote these words, been considerable progress towards the development of a cognitive process model. Jones' own careful and detailed observations provide invaluable raw material; Beghtol [5], in the course of an ingeniously argued model of the process of bibliographic classification, has given an account of progress in text linguistics in the field of summarisation. This area has been more fully described by Hutchins [6]. There have also been a number of developments, not directed specifically at the explication of the indexing process but relevant nevertheless, in cognitive psychology and particularly in reading research.

In this study an attempt will be made to outline a cognitive process model of document indexing on the basis of evidence from these disciplines. As well as indexing in the narrow sense, other relevant forms of document summarisation, namely abstracting and bibliographic classification, will also be considered.

Journal of Documentation, vol. 47, no. 2, June 1991, pp. 149–166

2. TASK DESCRIPTIONS

The starting point of any exercise in the cognitive modelling of a task is a detailed description of the task. A suitable task description is to be found in the relevant *British Standard* [7]. This, while necessarily a prescription rather than a description, is clearly based on the best current practice. The *Standard* identifies three stages in the indexing process:

1. Examining the document and establishing the subject content. The *Standard* states that 'ideally, full understanding of these documents depends upon an extensive reading of the texts'; but as this is regarded as being often impracticable, and not always necessary, a checklist is given of the parts of the text that need to be considered most carefully. Warnings are given against indexing from title alone or from an abstract. Other sources confirm this emphasis on rapid scanning. For indexing technical reports Cleverdon [8] recommends an optimum time of four minutes, which clearly allows for nothing more than scanning. For back-of-book indexes Anderson [9] states that books are frequently indexed in haste, and that the time allowed by a publisher is rarely enough to permit the indexer to read through the entire text. The British Standard recommendation for back-of-book indexes [10] has a similar message.

2. Identifying the principal concepts present in the subject. For the identification of relevant ideas to be indexed, the *Standard* recommends that 'agencies should establish a checklist of those factors which are recognised as important in the field covered by the index'. On the selection of concepts, there is a warning not to interpret too narrowly the breadth of interest covered by an index; so that, for example, when scientific and technical literature is being indexed its social or economic aspects should not be overlooked. Nevertheless, the indexer should attempt to choose the concepts that the community of users for whom the index was primarily designed would regard as most appropriate, and should take account of feedback from enquiries.

3. Expressing selected concepts in the terms of the indexing language. Most of this section consists of procedural advice on such matters as the checking of controlled descriptors. The *Standard* notes that indexers should be aware that a thesaurus or prescribed list of subject headings 'may not permit the exact representation of a subject encountered in a document'.

The *Standard* concludes with a section on quality control, emphasising the need for quality and consistency in indexing, and identifying factors affecting these: the impartiality of the indexer; his or her knowledge of the field covered by the documents to be indexed; the desirability that indexers should have direct contact with users; and the hospitality of the indexing language to new terms, changes in terminology, and new needs of users.

Finally, it may be noted that the *Standard* includes classification in the indexing process, and its recommendations for indexing apply equally to

bibliographical classification systems. For bibliographical classification specifically, a representative task description may be found in Mills and Broughton [11]. For abstracting, an unusually detailed description has been provided by Cremmins [12].

3. COMPREHENSION OF TEXT FOR INDEXING

The task description suggests that the requirements for a cognitive model of the indexing process fall into two groups, relating firstly to the indexer's comprehension of text, and secondly to the production of the index representation. It is reasonable to assume that indexers comprehend text in essentially the same way that fluent readers comprehend text, but with four modifications that are suggested by the task descriptions:

1. Indexers normally work under time constraints, which require them to scan text rapidly rather than read it at a normal reading rate.
2. Most indexers comprehend text solely for the purpose of classifying, indexing or abstracting the document containing that text. That task completed, the indexer has no further interest in the text. (Author-compiled back-of-book indexes and personal indexes are exceptions to this generalisation). These conditions are quite different from the conditions under which a reader's comprehension of text is measured in most psychological experiments, which typically seek to measure how much of a text a reader has understood or remembered after an interval.
3. The comprehension of text by indexers is followed directly by the production of an abstract, set of index entries, or classification. A model of the comprehension of text by indexers must therefore be directly linked to an appropriate production model.
4. Many indexers work within a narrow range of text types and subject fields, and the consequent repetitive element in their work leads to automatic processing beyond that associated with normal fluent reading.

There are many textbook accounts of the process of fluent reading, for example those by Mitchell [13], Just and Carpenter [14], and Smith [15]. This study will therefore concentrate on those aspects that are of particular relevance to the indexing process.

All these accounts agree that a reader's comprehension of text involves both top-down and bottom-up processing. Top-down processing uses information that is not contained in the text, but is part of the world-knowledge that the author possesses and which he assumes that his readers will also possess. This kind of processing is termed conceptual, to distinguish it from the perceptual or bottom-up processing of the information that is actually contained in the text. It is widely held [e.g. 15, pp. 13–16] that a person's world-knowledge is organised for storage in memory into large scale packages or structures known as frames, schemata, scripts, etc. Their precise organisation and nomenclature have been widely debated, but it is agreed that the appearance in a passage of text of one constituent of a frame/schema/script will bring to the reader's mind other constituents of the package and/or its overall topic.

When an indexer scans a document, what part of the scanning process is conceptual and what part is perceptual? What is the relationship between the two kinds of processing? While these questions have not been specifically researched, there is a large body of research evidence about speed reading generally, and about the relationship between reading speed and comprehension. These questions will now be addressed.

3.1 Speed reading and scanning

Speed-reading has a long history – Just and Carpenter cite papers going back to the first decade of the century – but almost until the past decade the practice has had an aura of thaumaturgy which has both led to its being exploited commercially and has also infected much research. Just and Carpenter claim [14, p. 428] that 'many studies of speed readers don't check what the speed reader learned, while others use poorly constructed tests, and still others fail to evaluate the speed reader's performance in comparison to the performance of an appropriate control group'. So we read of studies such as those of McLaughlin [16] whose rapid reader (3,500 wpm) read pages of 260 words in fourteen fixations (these are the pauses between saccades, or eye-movements), or of Thomas [17] who found someone capable of 10,000 wpm; this person made an average of six fixations per page, scanning down the left-hand page and up the right, and making no fixations at all on the bottom third of the page. Neither of these studies measured the extent of their subjects' comprehension of the text. One early experiment which did take account of comprehension was reported by Ehrlich [18], whose subjects, who had just completed a speed-reading course, were given a passage to speed-read in which a completely different text had been intercalated on every alternate line; none of them noticed. In a similar vein, Homa [19] assessed the perceptual and comprehension skills of two speed-readers who achieved between 15,000 and 30,000 wpm. He found their comprehension no better than that of normal students, and his cynical conclusion was that their only extraordinary talent was their ability to turn pages over quickly. The most thoroughgoing attempt to measure the comprehension of speed-readers was made by Carver [20]. He investigated the reading rate and comprehension of sixteen individuals who represented four categories of superior reader: people who had undergone a speed-reading course, professional persons, college students, and people who had scored highly on a test of reading speed. All scored 300–600 wpm for $>75\%$ comprehension, and speed readers were found to be similar in ability to other superior readers, except that they typically chose to skim at $>1{,}000$ wpm, and accept the lower comprehension that accompanies skimming.

The similarity between speed reading and scanning for indexing has been implied by Just and Carpenter. 'It is likely', they write [14, p. 429], 'that speed readers aim for a different type of comprehension than normal readers, a type that does not attend to details or to local coherence between ideas'. It is reasonable to suggest that this is precisely the kind of comprehension that indexers require, their purpose being to encapsulate comprehension in a handful of isolated keywords, details and local coherence being unnecessary distractions.

3.2 Conceptual and perceptual processing in scanning

Research carried out by Barrus, Brown and Inouye [21] might initially suggest that gist comprehension of the kind described above is sufficient even for abstracting. They studied the ability of speed readers to scan a chapter in a textbook and then to write an outline of the chapter. The outlines produced by speed readers working at 1,800 wpm were judged to be as good as those of normal readers working at 320 wpm. These results must however be treated with caution. The subjects were psychology students, and the chapter used was taken from a textbook of social psychology. One would expect therefore that much of the information obtained when skimming would have been processed conceptually rather than perceptually: that the subjects would have worked to a large extent by inference, fitting what their eye touched on into their own specialist knowledge.

The conceptual and perceptual aspects of scanning have been studied in detail by Masson [22] and by Just and Carpenter [14, pp. 429–449]. Masson cites previous research to show that skimmers process stories selectively, looking for information that is relevant to their purpose. This selective processing can take place at two levels. Perceptually skimmers look for cues in the text, e.g. long or italicised words, lead sentences in a paragraph, or using their knowledge of the text structure to fix their eye on goal-relevant areas of text – which is clearly identical to the text-processing methods of indexers and abstractors, as described by Jones [1, 23] and Cremmins [12]. The second level is conceptual. A sentence that has been selected for processing may be read completely or superficially. The reader may or may not make inferences to connect the constituent propositions [24], and may or may not use the information contained in the sentence to help produce a coherent representation of the whole text [25].

In a series of experiments investigating conceptual and perceptual processes in skimming stories, Masson [22] showed that readers when skimming were unable to distinguish perceptually between important and unimportant information in a story. Neither did readers appear to make use of the story structure to aid perceptually selective processing. However, readers were still able to process text conceptually in a way that allowed them to form accurate inferences about the story's macrostructure. Again, when given a specific goal in reading, selective processing of gist information was largely conceptual and not perceptual. Masson concluded by reviewing previous research relating to selective processing strategies that can be applied at the conceptual level. If an important statement is sampled, the reader may choose to read it carefully [26, 27]. Miller and Kintsch [28] showed that when reading gist-related information, a reader can select key propositions to help form a coherent macrostructure and speed the interpretation of newly processed information. Finally, a reader may draw plausible inferences to help connect propositions [25].

Just and Carpenter's experiments [14, pp. 429–449] were broadly similar in aim to Masson's, but were directed at the eye fixations and comprehension of trained speed readers – that is, college students who had recently completed a

commercial speed-reading course. They found that trained speed readers achieved a greater comprehension than untrained subjects reading at similar speeds, but that this was limited to comprehending the gist of texts on familiar topics; for answering questions of detail or questions on unfamiliar topics the trained speed-readers were no better than the untrained ones. Just and Carpenter's findings support those of Masson: namely that skill in speed-reading is conceptual rather than perceptual. Training in speed-reading did not teach readers to fixate on the most important words in a text; both trained and untrained readers were more likely to fixate on content words than on function words, though for trained readers this was due to the generally greater length of content words. In rapid reading, they conclude, text is sampled almost fortuitously, and the skill that trained speed-readers acquire is skill in inferring connections between the bits of text that they happen to have sampled. If the material is familiar, speed readers will possess schemata that are sufficiently detailed to support their inferences. These findings indicate the need for indexers to have specialist knowledge of the subjects they are indexing. Conversely, such schemata will be lacking if the material is unfamiliar. Relevant German research [29] is reported by van Dijk and Kintsch [30, p. 53]: students were given an essay to read and summarise, and 'appeared unable to distinguish what was macrorelevant in a text when they tried to summarise it right away after reading it once'.

3.3 Scanning of text for indexing and abstracting

A distinction that a comparison of speed-reading and indexing reveals is that between skimming and scanning text, in that scanning carries connotations of searching whereas skimming does not. There is ample evidence in the professional literature that indexers and abstractors scan text selectively, looking for specific perceptual cues. Some of these may be purely typographical – italicised words, headings, beginnings of paragraphs – but many are verbal. These verbal cues appear to fall into three groups.

1. *Word frequency.* It is practically a truism that the frequency of occurrence of a word or phrase is an important factor in determining whether to select it for indexing. We may infer then that an indexer, having noted a word – let us say 'carpets' – early on in a document as being a candidate for indexing, will have that word mentally foregrounded for the remainder of the scanning of the text, so that future occurrences of the word will be cued.

2. *Semantic nets.* Jones [23] has also argued that the indexer's interplay of perceptual and conceptual processing is such that, having cued 'carpets', he will also be more likely to cue semantically related words: 'floorcoverings' for example (which would in any case by conspicuous on account of its length), or 'rugs' (which would not). This has been tested experimentally: there is a sizeable body of psychological evidence on the priming of semantic networks, beginning with Collins and Quillian [31].

3. *Structural features.* There is evidence that abstractors at any rate carry in their heads a set of stock words and phrases pointing to structural features of a text: 'introduction', 'conclusions', 'In this paper we', 'Results suggest', etc. [12].

What light does this throw on the indexing process? The conclusion that indexing will be more effectively carried out if the indexers know something of the subject matter of their material is hardly surprising. Neither is the evidence from Masson that the selective processing of gist information was largely conceptual when subjects were given a specific goal in their reading: the indexing (or abstracting or classification) of documents is after all a highly specific goal. As so often happens, the value of this research is that it confirms the intuitions of professionals.

3.4 Cues in perceptual processing

Perceptual processing takes the form of scanning the text for cues: long words, words that are italicised or otherwise made prominent, lead sentences, and goal-relevant areas of text. These conclusions tie in well with Jones' observations about the text-scanning practices of indexers [1]: 'rare or unusual words or long words appear to be important'; 'if a word occurs in an opening paragraph ... this may heighten its suitability as an indexing term'; 'the opening sentences of individual paragraphs appear to be more significant'; and words that are defined in the text are likely to be chosen as indexing terms. On goal-relevant areas of text, the same author's earlier paper [23] noted that indexing operates at a number of levels, one of which is a 'structural or textual framework level', where it is claimed that authors jot down, or at least carry in their heads, 'skeletal structures' of what they are writing, and the indexer's task is to 'disinter this skeleton' by searching for surface clues.

The perceptual processing of text for abstracting is particularly concerned with goal-relevant areas of text. Cremmins, a professional abstractor, gives as the first two stages in the composition of an abstract: (1) 'Focussing on the basic features of the materials to be abstracted', i.e. their form: monograph, article, dissertation, etc.; their type: experimental research, survey, description, review, etc.; how the text is structured: whether primary and secondary headings are used, 'particularly those containing such guide words as "introduction", "methods", "results", "conclusions", and recommendations; if there are conclusions whether they are presented together or scattered throughout the text', etc. (2) Identifying relevant information: rapid-reading the text to identify cue words: 'In this paper we', 'Administration of', 'Data were analyzed', 'Results suggest', etc; or concentrating on 'information presented under conventional functional headings such as "Introduction" and "Methods" '; or checking the first and last sentences in a paragraph which 'often are topical or summary ones' [12, p. 15–18].

It is clear then that professional indexers (and especially abstractors) develop an awareness of the structural properties that are inherent in text irrespective of its subject content. These properties fall into two groups. Firstly the overall structure of the text must be considered. In recent years

considerable theoretical and experimental work has been carried out into overall text structures. Hutchins [32] has described and consolidated much of this research, and further elaboration is provided by Hoey [33]. Secondly, some parts of a text are more significant than others for the purpose of providing information for abstracting or indexing: Anderson [9] has anticipated Jones in giving as examples opening paragraphs of chapters, sections etc., and the opening sentences of paragraphs.

The question that must now be asked is: given that any discourse has a structure that conforms to one (or perhaps a combination) of a limited number of patterns, how is an indexer/abstractor to recognise this? Under normal conditions of reading a person is able to make use of a wide range of lexical devices that aid comprehension. These are variously described in the literature of psycholinguistics, but are reducible to two: *connectives*, words and phrases that relate clauses and sentences, and the making of *inferences* 'based on the reader's knowledge of the world, of language, and of the text portions that have already been read' [14, p. 252]. These accounts presume however that the reader is able to adopt a normal reading rate; and that as we have seen is a luxury that the working indexer or abstractor rarely enjoys. Our task is therefore to investigate the perceptual cues that are available under conditions of rapid scanning. Both Cremmins [12] and Jones [23] have given detailed accounts of perceptual cues from their professional points of view. Cues for abstracting are not the same as those for indexing: the cues that Cremmins the abstractor describes are structural, whereas Jones' cues for indexing are topical, their purpose being to help the indexer search for keywords. Research evidence attests to the soundness of Jones' assertions. Just and Carpenter [14, p. 46–47] review the research, much of it their own, which demonstrates that a reader's gaze duration increases linearly according to the number of letters in a word. Jones' advice to look for unusual letter combinations echoes long-standing controversies over whether the word-encoding process takes as its input the overall shape of a word, or whether words are identified letter by letter, or through the recognition of clusters of letters [15, p. 46-47]. The use of typographical cues has been researched by Glynn and Di Vesta [34]; and Hartley and Trueman [35] have reviewed the literature on the effect of headings on recall, search and retrieval.

For Jones' final cue-type, 'the distinctive syntax of definitions', the evidence may be more equivocal, but may at the same time offer an insight into the mental processes of indexing. The importance of definitions in expository text is clear [36]; what is debatable is the extent to which the syntax of definitions provides an adequate cueing mechanism when text is being scanned rapidly. A common syntactic type is, to quote Jones [23], 'An x is a y'. It is difficult to imagine a more minuscule perceptual cue than 'is a' to fixate on when scanning a page of text rapidly. If so insignificant a phrase can indeed be used as a perceptual cue, it can only be because indexers have trained themselves to recognise it. Even then, it may be argued that two other devices reinforce its perceptual significance. One is an unfamiliar word: the syntactic pattern 'An x is a y' expands to 'An [unfamiliar word] is a [phrase defining the unfamiliar

word]', the unfamiliarity of the word providing an additional perceptual cue. The other device is structural, in that there is according to Jones a likelihood that definitions may appear in the first sentence of a paragraph. If this is the case, then the phrase 'is a' illustrates for all its insignificance the complexity of the indexer's expertise in bringing into simultaneous play three different types of perceptual cue when scanning a page.

3.5 Conceptual knowledge and the development of expertise

Because of the importance of conceptual processing, it is worth pausing to consider the range of concepts that indexers keep in their memory. Two kinds have been identified in the discussion so far, relating to (1) the subject matter to be indexed, and (2) the structure of the texts to be scanned. Three more may be suggested. Two are axiomatic in professional terms and are explicit in the task description: indexers, abstractors and classifiers alike need to know about (3) the systems they are using, and about (4) the users of those systems. Finally, (5) a background of general world-knowledge is necessary for the comprehension of any spoken or written discourse [see e.g. 15, pp. 6–22].

An indexer's conceptual knowledge will clearly take time to develop: Cremmins has remarked on the contrast between the working speeds of novice and expert abstractors. As well as being slow, novices are more likely to suffer from a lack of understanding of their subject matter. There are indications that this may follow a consistent pattern. Dee-Lucas and Larkin [37] have studied the comprehension of physics textbooks by physics students at different stages, and found that beginning-level students 'consider certain information to be important simply because of its category membership (i.e. whether it is a definition, equation, fact), regardless of its content. . . . Novices consider the same substantive information to be more important when presented as a definition (rather than a fact) and as an equation (rather than a verbal phrase)'. It is just possible that indexers, even expert ones, may be subject to similar biases: Jones [23] stressed the importance to the indexer of definitions and of unusual typographical effects, of which an equation would be an example. On the other hand it is arguable that watching out for definitions is just one of a range of procedures that comprise an experienced indexer's armoury of techniques; according to Charney and Reder [38] 'having a skill means knowing when to apply particular procedures'. Dreyfus and Dreyfus [39] similarly argue that beginners in a task follow rules mechanically, without any coherent sense of the overall task. A proficient performer, on the other hand, will understand 'without conscious effort what is going on, [but] will still have to think about what to do'. A real expert – Dreyfus and Dreyfus postulate five levels of expertise, of which this is the fifth and highest – 'understands, acts and learns from results without any conscious awareness of the process'.

If skilled indexers do have the ability to distinguish what is macrorelevant in a text after not so much reading it once as scanning it once, then the ability must come by dint of long practice. Numerous studies attest to the development of skilled memory – 'the rapid and efficient utilization of memory

in some knowledge domain to perform a task at an expert level' [40]. This paper by Chase and Ericsson introduces one of the most frequently cited of all studies of skilled memory acquisition: the student who, after 250 hours of practice spread over two years, was able to recall a string of eighty-one random digits.

Skilled memory was found to be based on three principles: (1) the use of meaningful associations to encode knowledge into long term memory (LTM) in the form of structured groups (which are often called chunks); (2) the use of retrieval cues that are explicitly associated with the memory encoding, so that retrieval from LTM is triggered by means of these cues; and (3) the importance of practice in increasing the speed of encoding and retrieval operations. Ericsson and Oliver [41] report that these principles have been found to apply to other subjects performing similar memory tasks. For indexing and abstracting activities specifically, however, the evidence is less certain, though it is possible to identify some areas where these principles apply: for example the ability of skilled indexers and abstractors to retrieve automatically from memory a wide range of structural and topical cues. There have been numerous studies of the development of complex cognitive processing skills, and a review by Colley and Beech [42] has been published.

4. A BASIS FOR A COGNITIVE PROCESS MODEL

It was shown in section 3 above that the indexing process has two stages: the comprehension of text for indexing, and the production of the index terms. An indexer's comprehension of text was seen to have only a limited number of defined differences from that of a normal fluent reader. The construction of a process model of indexing will be greatly simplified if it can be based on an existing model of reading comprehension that is capable of accommodating these differences.

Discourse comprehension is a much-researched topic, and a number of models have been proposed in recent years: those by Frederiksen [43], Meyer [44], Kintsch and van Dijk [25], van Dijk and Kintsch [30], and Johnson-Laird [45] may be cited. The best known model is perhaps that of Kintsch and van Dijk, which incorporates the model of text reduction proposed by van Dijk [46]. This offers a simple and coherent set of procedures (called macrorules) for reducing text to its gist elements. However, as Beghtol [5] has observed, the applicability of van Dijk's model to the indexing process is problematical, since in the indexing situation conditions of rapid scanning are often inconsistent with the kind of close reading that the macrorules require. The latest (1983) version of van Dijk and Kintsch's model [30] goes some way towards meeting this objection and is based on processes rather than on rules. The earlier macrorules are replaced with macrostrategies, which

> are flexible and have a heuristic character. A language user need not wait until the end of a paragraph, chapter, or whole discourse before being able to infer what the text or the text fragment is about, globally

> speaking. In other words it is plausible that with a minimum of text information from the first propositions, the language user will make guesses about such a topic. These guesses will be sustained by various kinds of information, such as titles, thematic words, thematic first sentences, knowledge about possible ensuing global events or actions, and information from the context . . . An expedient strategy will operate on many kinds of information, which individually are incomplete or insufficient to make the relevant hypothetical assumption [30, pp. 15–16].

The process of reducing text to its gist elements (i.e. the inferring of macrostructures) consists of the application of a range of contextual and textual macrostrategies. Contextual macrostrategies concern conceptual processing: the reader limits semantic searches to the general cultural context of the writer; the reader decides which topics are characteristic of the discourse type expected in a particular context; and so on. Textual macrostrategies involve the considerations of text structure and perceptual cues that have already been discussed in the present study. For our purposes the 1983 macrostrategies have the advantage over the earlier macrorules of being less dependent on bottom-up processing: they admit the processing of text and the inferring of a macrostructure without the need for the reader to have read every word. If the earlier macrorules have a place in the modelling of the process of indexing, it will be as a backup operating locally, where the indexer feels the need to pause and read carefully through a particular passage.

Van Dijk and Kintsch's 1983 model also has the advantage of adaptability, as its authors intended it to be used as a framework that could be adapted to different situations. Because of its hospitality to rapid scanning, the model adapts quite easily to the indexing process, and will form the basis for the model to be proposed here.

5. COMPREHENSION MODEL

Van Dijk and Kintsch's 1983 model uses the three conventional divisions of memory: a sensory register, short-term (working) memory, and long-term memory.

5.1 Sensory register: this is 'a theoretical necessity rather than a known part of the brain' [15, p. 89]. Its function is to convey visual information to working memory. The principal source of this information is, quite obviously, the document being worked on. There will in many cases be additional information taken from controlled language schedules, reference sources, and the like, or from consultations with colleagues.

5.2 Short-term memory (STM), or working memory, is defined as having both processing and limited storage functions. In a classic paper Miller [47] defined STM as having primarily a storage function. (A tempting instance of his 'seven plus or minus two' formula occurs in Campbell's survey [48] which found that

indexers assigned on average 7.6 index terms to a document.) Later research has modified Miller's theory by assigning processing as well as storage functions to STM: Broadbent [49] suggested that a maximum of three or four items can be processed simultaneously, which fits our suggestion that the scanning of a page of text for indexing involves the simultaneous processing in STM of three types of perceptual cue.

5.3 Long term memory (LTM), comprises 'the totality of an individual's knowledge and beliefs about the world' [15, p. 310]. In van Dijk and Kintsch's model LTM has three interacting divisions:

5.3.1 Episodic text memory. The metaphor in van Dijk and Kintsch is of push-down stacks, to which new elements are constantly being added, pushing the older elements further and further away. This may be an appropriate image for a general model of textual comprehension, though current thinking tends to favour a more connectionist approach. For this reason, and because the tasks of classification, indexing and abstracting all require a limited and functional comprehension, it may be more appropriate to visualise episodic text memory as consisting of networks of interrelationships.

5.3.2 Relevant knowledge, which for our purpose consists of the five categories of concepts that were identified in section 3.5 above.

5.3.3 A control system, containing elements that are brought unconsciously into play as a background influence on the way in which processing is carried out. The concept of a control system as stated by van Dijk and Kintsch comprises 'the comprehender's goals and purposes, wishes, interests and emotions'. Some of these are transient qualities, and for present purposes they may need to be supplemented by the more durable elements of cognitive ability and cognitive style, in order to provide a fuller explanation of individual differences in processing.

The starting point of processing may be described as a *situation model.* One way to define a situation model is to regard it as containing the indexer's initial impression of a document – normally its title. Alternatively, a situation model might be regarded as containing the indexer's preconception of the kind of document he is about to index, which would give a far more generalised situation model, and one residing in LTM rather than in working memory. The model would incline to one or the other version depending on the breadth of material handled by the indexer. Either way, further processing (i.e. scanning) will result in the crystallisation of the situation model into an *aboutness model*, where the overall aboutness of the document has been identified.

In van Dijk and Kintsch's framework model, the processing that is carried out in working memory is principally the formation of propositions and their selective chunking or rejection to produce macropropositions that are constantly being adapted and despatched to and recalled from long-term

memory. In indexing, processing along those lines takes place locally and selectively, when the indexer slows down the scanning in order to read passages that have been identified as being particularly significant. The identification of significant passages is achieved through the selective acceptance of visual cues from the sensory register; and for this purpose working memory is defined as having a limited storage capacity, acting as a buffer in which cues are kept while analogies are being sought out and fetched from LTM.

There is constant interaction between working memory and LTM as part of the continuous process of amplification and adjustment from the initial situation model to the final goal. With experienced indexers much of this interaction is automatic. The visual cues that enter STM from the sensory register interface initially with the indexer's relevant knowledge and control system, i.e. with the indexer's preconception of what the document should be about and with the goal and purpose of the processing. As processing continues, further cues pass between STM and episodic text memory as candidate index terms. The operations carried out in STM are: (*a*) *Reinforcement*: a candidate index term is consistent with the overall aboutness of the document and/or matches one already processed. (*b*) *Modification*: a candidate index term is modified in conformity with the indexer's relevant knowledge and/or goal, or with a candidate index term that is already in episodic text memory. (*c*) *Chunking*: a number of semantically related candidate index terms are merged to form a single term. (*d*) *Rejection*: a candidate index term is inconsistent with the indexer's aboutness model, relevant knowledge, or goal.

6. PRODUCTION MODEL

Whereas models of discourse comprehension are readily found, models of discourse production are far less common. Van Dijk [50] explains this as being due to the comparative lack of initial data: for a comprehension model, the discourse is the initial data, which can be tested for comprehension, storage and so on. For a production system on the other hand the initial data consist of vague and unspecified 'ideas', 'wants', etc., which are far more difficult to handle under experimental conditions. It is perhaps fortunate for our present purpose that indexing, abstracting and classification require production models that are simple, highly structured, and directly derived from a comprehension model. These productions are, simplest first:

1. A classification production, which in bibliographic classification generates a single coded representation of the overall aboutness of the document.

2. An indexing production, which generates one or (usually) more index terms which point to (indicate) themes contained in the document.

3. An abstracting production, which generates a prose statement of the document's aboutness, and in the case of an informative abstract, a statement

of the methods and conclusions of any investigations described in the document.

All three productions are subject to constraints imposed by the information systems of which they form part:

1. Classifications are invariably and necessarily 'closed' systems, which permit the classifier to specify only such topics as the system has made provision for, and impose a fixed pattern on the structure and sequencing of topics. To take a typical example, Austin [51, p. 283] describes a work whose overall aboutness is represented in the Dewey Decimal Classification by a notational symbol, 598.29410222, denoting 'birds', 'British', and 'illustrations' – in that order; the system does not permit the ordering of these topics to be altered in any way. In addition the Dewey system imposes particular hierarchical structures (e.g. zoology – vertebrates – birds) on the topics. Finally, the work that Austin used for his example was one describing specifically how bird tables can be used to attract birds to gardens; the Dewey system makes no provision for gardens and bird tables to be represented in this context, and they have accordingly been omitted from the specification.

2. In indexing, the extent of the constraint varies according to whether the system uses a controlled, natural or free language. In controlled language indexing systems the constraints are similar to those in classification (except that there are fewer built-in hierarchical structures). Natural language systems are limited to using terms found in the work; in the case of free language systems even this constraint is removed. An environmental constraint to which indexing systems of all types are subject is that of exhaustivity.

3. In the case of abstracting, the constraints are all environmental, relating to editorial policy on abstract type, length, literary style, and readership orientation.

While it is convenient to describe the comprehension and production models separately, in practice the two merge into one another, in that a candidate index term is an element in both the indexer's comprehension of the text and in the resulting index production. It is obvious too that production requirements will influence comprehension: the more exhaustive the index, the more detailed the comprehension that will be required. If the specific production requirement is classification or an abstract rather than a set of index terms, the comprehension requirements will differ again. The common element on which all these productions is based is the aboutness model, i.e. the point at which the indexer, classifier or abstractor has determined the overall topic of the document. It is after this point that processing follows different routes according to whether the document is being classified, indexed, or abstracted.

1. For classification, the aboutness model is compared with the schedules of the classification system. Modifications to the aboutness model will be made in cases where the classification system is unable to accommodate the aboutness model intact. This may involve the identification of keywords, particularly if the classification is to be used as the basis for indexing decisions.

2. For indexing, the aboutness model will in every case lead to the identification of keywords, or candidate indexing terms. In controlled language indexing these will be compared with a schedule – a thesaurus, list of subject headings, or authority file of previous decisions – and modifications made where the candidate term does not exactly match the controlled language descriptor.

3. For abstracting, the aboutness model is followed by rather more detailed processing than is the case with classification or indexing. The processor is required to identify both structural cues – identifying for example whether conclusions are presented in the text – and subject cues; and these must be further processed to form a statement in continuous prose.

Perceptual processing of text will normally end before the achievement of the goal-state. In classification, it ends as soon as the aboutness model is compared with the schedule; in indexing and particularly in abstracting processing of text is likely to tail off rather than cease abruptly. Loops may be inserted into the processing at any point after the initial situation model and before the final attainment of the goal. These may be of any number and length, from regressive eye-fixations at saccadic level to the complete re-scanning of the text.

DISCUSSION

The model presented here is just one of many possible ways of modelling the indexing process. It is based on a varied body of research evidence, but that research has mostly been directed more towards the comprehension of text than towards any specific production goal. Even though the model is in many ways sketchy and incomplete, some clear principles emerge: the importance of conceptual processing and the consequent need for indexers to be familiar with the subject matter of the texts they are working with; the limited comprehension obtainable by scanning and the types of perceptual cues in scanning text and their use; and the nature and importance of expertise in the performance of a professional task. More specifically, in the education and training of indexers, the model helps to pinpoint the causes of inadequate or inaccurate indexing. On a psychological level, the model has implications for the design and implementation of information retrieval systems, particularly in drawing attention to the need for systems to take into account the limited storage and processing capabilities of working memory. It is possible too that the model could point towards the cognitive simulation of the indexer's

decision-making processes. It is likely that more implications would emerge from research. There is in any case a need for more research into the indexing process itself, if only because the validity of the analogies and parallels contained in the model is untested. It is surely remarkable that so little is really known about so basic a professional activity.

REFERENCES

1. JONES, K.P. How do we index? a report of some Aslib Information Group activity. *Journal of Documentation, 39*, 1983, 1–23.
2. COLLISON, R. *Indexing books: a manual of basic principles.* London: Benn, 1962.
3. KNIGHT, G.N. *Indexing, the art of: a guide to the indexing of books and periodicals.* London: Allen & Unwin, 1979.
4. BORKO, H. *and* BERNIER, C.L. *Abstracting concepts and methods.* New York: Academic Press, 1975.
5. BEGHTOL, C. Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation, 42*, 1986, 84–113.
6. HUTCHINS, J. Summarization: some problems and methods. *In*: JONES, K.P., *ed. Informatics 9: Meaning: the frontier of informatics: proceedings of a conference.* London: Aslib, 1987.
7. BS 6529.1984. *British Standard recommendations for examining documents, determining their subjects and selecting indexing terms.* London: British Standards Institution, 1984.
8. CLEVERDON, C.W. *Aslib Cranfield research project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems.* Cranfield, 1962.
9. ANDERSON, M.D. *Book indexing.* Cambridge: Cambridge University Press, 1971.
10. BS 3700:1988 *British Standard recommendations for preparing indexes to books, periodicals and other documents.* London: British Standards Institution, 1988.
11. MILLS, J. *and* BROUGHTON, V. *Bliss bibliographic classification. 2nd ed. Introduction and auxiliary schedules.* London: Butterworths, 1977.
12. CREMMINS, E.T. *The art of abstracting.* Philadelphia: ISI Press, 1982.
13. MITCHELL, D.C. *The process of reading: a cognitive analysis of fluent reading and learning to read.* Wiley, 1982.
14. JUST, M.A. *and* CARPENTER, P.A. *The psychology of reading and language comprehension.* Boston: Allyn and Bacon, 1987.
15. SMITH, F. *Understanding reading: a psychological analysis of reading and learning to read.* 4th ed. Hillsdale, NJ : Erlbaum, 1988.
16. McLAUGHLIN, G H. Reading at 'impossible' speeds. *Journal of Reading, 12*, 1969, 449–454, 502–510.
17. THOMAS, E.L. Eye movements in speed reading. *In*: STAUFFER, R.G., *ed. Speed reading: practices and procedures.* Newark, NJ : University of Delaware, Reading Study Center, 1962.
18. EHRLICH, E. Opinions differ on speed reading. *National Education Association Journal, 52*, 1963, 45–46.
19. HOMA, D. An assessment of two extraordinary speed-readers. *Bulletin of the Psychonomic Society, 21*, 1983, 123–126.
20. CARVER, R.P. How good are some of the world's best readers? *Reading Research Quarterly, 20*, 1985, 389–419.
21. BARRUS, K., BROWN, B.L. *and* INOUYE, D. *Rapid reading reconsidered.* Paper presented at the meeting of the Psychonomic Society, San Antonio, Tx, 1978.

22. MASSON, M.E.J. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8*, 1982, 400–417.
23. JONES, K.P. Towards a theory of indexing. *Journal of Documentation, 32*, 1976, 118–125.
24. BOWER, G.H., BLACK, J.B. *and* TURNER, T.J. Scripts in memory for text. *Cognitive Psychology, 11*, 1979, 177–220.
25. KINTSCH, W.A. *and* VAN DIJK, T.A. Towards a model of text comprehension and production. *Psychological Review, 35*, 1978, 363–394.
26. CIRILO, R.K. *and* FOSS, D.J. Text structure and reading time for sentences. *Journal of Verbal Learning and Verbal Behavior, 19*, 1980, 96–109.
27. JUST, M.A. *and* CARPENTER, P.A. A theory of reading: from eye-fixations to comprehension. *Psychological Review, 87*, 1980, 329–354.
28. MILLER, J.R. *and* KINTSCH, W. Readability and recall of short prose passages: a theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 1980, 335–354.
29. SCHNOTZ, W., BALLSTAEDT, S. *and* MANDL, H. Kognitive Prozesse beim Zusammenfassen von Lehrtexten. *In*: MANDL, H., *ed. Zur Psychologie der Textverarbeitung*. Munich: Urban & Schwarzenberg, 1981.
30. VAN DIJK, T.A. *and* KINTSCH, W. *Strategies of discourse comprehension*. New York: Academic Press, 1983, 53.
31. COLLINS, A.M. *and* QUILLIAN, M R. Retrieval time from semantic memory. *Journal of Verbal Language and Verbal Behavior, 8*, 1969, 240–247.
32. HUTCHINS, J. On the structure of scientific texts. *UEA Papers in Linguistics, 5*, 1976, 18–39.
33. HOEY, M. *On the surface of discourse*. London: Allen & Unwin, 1983.
34. GLYNN, S.M. *and* DI VESTA, F.J. Control of prose processing via instructional and typographical cues. *Journal of Educational Psychology, 71*, 1979, 595–603.
35. HARTLEY, J. *and* TRUEMAN, M. The effect of headings in text on recall, search and retrieval. *British Journal of Educational Psychology, 53*, 1983, 205–214.
36. KINTSCH, W.A. Text representations. *In*: OTTO, W. *and* WHITE, S , *eds. Reading expository material*. New York: Academic Press, 1982, 87–102.
37. DEE-LUCAS, D. *and* LARKIN, J.H. Novice rules for assessing importance in scientific texts. *Journal of Memory and Language, 27*, 288–308.
38. CHARNEY, D.H. *and* REDER, L.M. Initial skill learning: an analysis of how elaborations facilitate the three components. *In*: MORRIS, P., *ed. Modelling cognition*. Chichester: Wiley, 1987, 135–165.
39. DREYFUS, H. *and* DREYFUS, S. Why skills cannot be represented by rules. *In*: SHARKEY, N.E., *ed. Advances in cognitive science 1*. Chichester: Ellis Horwood, 1986, 315–335.
40. CHASE, W.G. *and* ERICSSON, K.A. Skilled memory. *In*: ANDERSON, J.R., *ed. Cognitive skills and their acquisition*. Hillsdale, NJ : Erlbaum, 1981, 141–189.
41. ERICSSON, K.A. *and* OLIVER, W.L. A methodology for assessing the detailed structure of memory skills. *In*: COLLEY, A.M. *and* BEECH, J.R., *eds. Acquisition and performance of cognitive skills*. Chichester: Wiley, 1989, 193–215.
42. COLLEY, A.M. *and* BEECH, J.R. Acquiring and performing cognitive skills. *In*: COLLEY, A.M. *and* BEECH, J.R , *eds. Acquisition and performance of cognitive skills*. Chichester: Wiley, 1989, 1–16.
43. FREDERIKSEN, C.H. Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology, 7*, 1975, 371–458.
44. MEYER, B.J.F. *The organization of prose and its effect on memory*. Amsterdam: North Holland, 1975.
45. JOHNSON-LAIRD, P.N. *Mental models*. Cambridge, Mass.: Harvard University Press, 1983.

46. VAN DIJK, T.A. *Text and context: explorations in the semantics and pragmatics of discourse*. Harlow: Longmans, 1977.
47. MILLER, G.A. The magical number 7, plus or minus two. *Psychological Review, 63*, 1956, 81–97.
48. CAMPBELL, D.J. *A survey of British practice in coordinate indexing in information/ library units*. London: Aslib, 1975.
49. BROADBENT, D.E. The magic number seven after 15 years. *In*: KENNEDY, R.A. *and* WILKES, A., *eds. Studies in long-term memory*. New York: Wiley, 1975.
50. VAN DIJK, T.A. *Macrostructures: an interdisciplinary study of global structures in discourse, interaction and cognition*. Hillsdale, NJ : Erlbaum, 1980, 282.
51. AUSTIN, D. *PRECIS: a manual of concept analysis and subject indexing*. 2nd ed. London: British Library Bibliographic Services Division, 1984.

(*Revised version received 30 January, 1991*)