All in the mind: concept analysis in indexing

John Farrow

The indexing process consists of the comprehension of the document to be indexed, followed by the production of a set of index terms. Differences between academic indexing and back-of-book indexing are discussed. Text comprehension is a branch of human information processing, and it is argued that the model of text comprehension and production developed by van Dijk and Kintsch can form the basis for a cognitive process model of indexing. Strategies for testing such a model are suggested.

Introduction

Ross Todd wrote in 1992 of the "urgent need for the profession to develop a theoretical knowledge base of indexing"; more recently the President of the Society of Indexers reiterated this view in a guest editorial. The emergence in the mid-1970s of cognitive psychology as a distinct science has led to the development of a body of research that can be applied to help us understand the indexing process. This paper outlines some of the features of this research, and tries to suggest how it might be used as a framework for empirical research into this process.

A number of researches have had as their aim the construction of a comprehensive theory of language understanding and text processing. When we read a passage of text, we can do so for one or more of a range of purposes. Our purpose may be affective: we may read for inspiration or amusement. It may be learning-orientated; by far the largest amount of research into human text processing has been in this area, as it is a lot easier to measure how much of a passage of text a person has remembered than to measure the amount of inspiration or amusement a person has derived from it. Indexers, however, do not in their professional capacity read either for amusement or to learn (though these may take place incidentally). Indexers' reading is task-orientated: the book or other document that an indexer reads is read for the specific and immediate purpose of producing an index. The indexer's comprehension of a document is accordingly linked to a production task based on that same text.

Similar comprehension-production processes govern the work of classifiers and abstractors, and of others too, such as the compilers of abridgements and synopses: their common process is content analysis and their common purpose is document representation. There are, however, differences in purpose which impact upon the content analysis. For a classifier, the purpose is a representation of the overall aboutness of the document, translated into the language of the classification system. For an abstractor, the purpose is to represent in continuous prose the document's principal arguments. Indexing is of two kinds. Back-of-book indexes enable readers to locate information about a topic within a book. The book's overall aboutness is already known, from such sources as the title and table of contents. The book indexer's task is to read the text, distinguishing between relevant and peripheral information, and employing much the same mix of top-down (conceptual) and bottom-up (perceptual) processing that takes place in normal fluent reading. "Academic indexing"—a useful shorthand term that has crept in by stealth in recent years to label the kind of indexing practised by indexing and abstracting databases and in library catalogues—uses a predominantly top-down approach. Here the indexing is far less exhaustive than is the case with back-of-book indexes. Many of the machine-searchable bibliographic databases commonly found today contain abstracts that are keyword searchable, and an increasing number carry the full text, similarly searchable. In machine retrieval, academic indexing supplements these fuller sources of searchable terms. If academic indexing is to have any purpose (a number of databases do not consider it necessary) it is to identify the principal topics of the document, at a level of exhaustivity somewhere between the title and the abstract.

Human information processing

In cognitive terms, the reading process is firmly located within the area of human information processing. The predominant approach to human information processing is summarized as follows. Information from the outside world is input by the senses into a sensory register or store. This holds information in an unprocessed form for a very brief period—perhaps around one quarter of a second. During this time the
information may be selected for further processing, either because the mind recognizes that the information constitutes a pattern which has been previously stored in memory, or because it wishes to select for further processing information that does not match any existing pattern. Information that is not selected at this stage is lost.

Selected information passes into short-term memory (STM). STM is held to be transient unless one or more of a number of control strategies are used to retain it in STM or transfer it to long-term memory (LTM). The capacity of STM is limited to ‘seven plus or minus two’ items. Items between which some structure, meaning or relationship can be identified are grouped by the mind into chunks, which are assimilated as units of information in their own right. For example, telephone numbers often contain up to eleven digits, arranged into groups of three or four digits which may be committed to memory as three chunks (area code, district code, subscriber number) and may be reduced further to a single chunk (telephone number). An implication of chunking is the constant transference of information between STM and LTM. The concept of working memory is therefore sometimes postulated as being supplementary to, or even synonymous with, STM, and acting as a kind of scratch pad against which items of information, which normally reside in LTM in an inactive state, are placed in an active state for as long as they are being worked upon.

Unlike STM, LTM stores information relatively permanently. Information is transferred between STM and LTM by means of control strategies which govern acquisition and retrieval. Acquisition strategies include rehearsal (the repetition of information, whether intentionally or unintentionally), coding (the use of an easily retrieved mnemonic), and imaging (the creation of visual images by which to remember verbal information); teachers everywhere will recognize these as basic tools of their trade. Retrieval strategies are not always easy to describe, as much retrieval occurs too rapidly to permit effective study. Categories and hierarchies of categories are an obvious way of organizing knowledge for retrieval, as structured information is more easily recalled than unstructured. The size of a knowledge group or category appears to range between two and five items, often three. Information may be organized into semantic networks. When a concept is activated, related concepts within the same network are also activated, the degree of activation depending on the semantic distance from the activated concept.

More complex knowledge domains are often represented by a range of associated structures variously called schemata, frames, and scripts, by means of which generic information about situations, events, objects—all that makes up our world—is stored in memory as knowledge clusters and used to recognize and understand new examples of the schema. When a schema is activated, new information is processed against it in order to establish the goodness of the match. Schema and frame are often used interchangeably, though specifically the term ‘frame’ is applied to knowledge-based computer systems. A script is a schema for a series of events—‘going to a restaurant’ is the oft-quoted example—with elements such as roles (diner, waiter), entry conditions (hungry, able to pay, restaurant open), and props (tables, chairs, food, etc.).

Text processing

This, then, is a very broad summary of human information processing. It has been used as a basis for a number of models of text processing, notably that developed by the Dutch linguist Teun van Dijk in conjunction with the American psychologist Walter Kintsch. A text, according to the model, consists of a coherent sequence of propositions, a proposition being the smallest unit of discourse that is capable of being evaluated. For example, the statement ‘Paul is writing a letter’ contains two propositions, to the effect that (1) Paul is writing, and (2) what he is writing is a letter.

The starting point of the model is the ability of language users to read a passage of text and sum up its theme, or topic, or what it is about, in an intuitively and conventionally acceptable sentence. The specific status of the sentences in a passage of text is to be sought in the relationship between their meaning and the meaning of the other sentences in the sequence. The meaning of these sentences is related inferentially to our knowledge of the world. Groups of propositions form a subsystem (schema) of knowledge about some phenomenon in the world. Relations between the propositions that make up a topic may be represented hierarchically, while at a higher level, a complex topic will be made up of subtopics. New topics are introduced, and the reader is made aware of their introduction, whenever one of the sentences of a discourse introduces an argument or a predicate which cannot be subsumed under higher order arguments or predicates of the given topic. Thus the aboutness of a discourse is hierarchical, and a text has macrostructures—overall semantic structures—at different hierarchical levels. The top macrostructural level often corresponds with the title of a text.

In comprehending a text, the reader links together the propositions that comprise a coherent sequence. Certain propositions are retained (using the strategies described above), and may be linked by chunking to the next sequence of propositions. Topic sentences (e.g. titles, section headings, first sentences of paragraphs) help the formation of macrostructures. If there is an insufficient argument overlap, search and inference operations involving LTM are required—in other words, a passage of text is difficult to understand. The process continues cyclically until it comprehends the whole text base.
If we are to obtain macrostructures of any sequence of text we must apply various operations of semantic information reduction. Besides being an essential element in text comprehension generally, macrostructures provide the indexer with the basis for deciding what parts of the text to index. Four rules (macrorules) govern information reduction (there is in addition a general rule governing conditions under which reduction may not take place). The macrorules are:

1. Weak deletion: information is simply left out. Only accidental information may be deleted: that is, information that may be omitted without changing the meaning or influencing the interpretation of the subsequent sentences of the discourse. Thus Law⁵, describing her indexing of Chambers’s Encyclopaedia, states: ‘On the problem of what references to take out, the main criterion was whether the reference conveyed material information about the subject of the actual index heading: allusions incidental to some other subject had to be passed over.’

2. Strong deletion: the deletion of locally-relevant information. Here, the information which is deleted specifies normal or expected associations; as when the indexer of Charles Dickens’ letters⁶ tells us: ‘A biographer of Dickens once told me that he would expect an index to show the number of times Dickens dined with Forster. This my index will show. It does not however say whether they dined at home, at a pub or a club, or the house of a mutual friend.’

3. Generalization: where several objects or properties of the same superordinate class are referred to, globally, with the name of the superordinate class. Bell⁷ has noted that when indexing biographies it is often necessary to index ‘abstract concepts not named on the page, such as “Courage”, “Loyalty”, “Fear”, “Patriotism”.’ Another example is to be found in G. N. Knight’s account⁸ of how he indexed the life of Winston Churchill: ‘In the case of CHURCHILL, WINSTON LEONARD SPENCER... what I have done... is to limit the full-blown subheadings to CHARACTERISTICS, FINANCES, HEALTH, HOBBIES and POLITICAL INTERESTS.’

4. Construction, or the combining or integration of information which denotes essential properties, causes, components, consequences, etc., of a higher-level fact. This constitutes what Mulvany⁹ calls ‘implicit information’, an example of which is found in a discussion of the mineral and vitamin content of dog food in a book about raising dogs. Although many types of minerals and vitamins are discussed in relation to wheat-based and corn-based dog food, never is the word “nutrition” used. In a case like this, “nutrition” is an implicit concept that will be identified by the indexer and added to the index.

This account of van Dijk and Kintsch’s model might at first sight appear to emphasize the perceptual (bottom-up) aspects of text processing. However, the model’s references to frames and topic sentences indicate that conceptual (top-down) processing plays at least as important a role. The information presented by the text is constantly being compared with knowledge stored in LTM; schema theory provides a powerful model for the organization of that knowledge.

Prior knowledge is used in text processing in a variety of ways. It is a basic component of the reading process, at all levels from letter recognition upwards. Specific background knowledge provides a context for abstract and unfamiliar statements, by enabling readers to relate what they are reading to what they already know. Broadly, people who already know something about a subject find it easier to learn more about that subject than people who know little. This underlines, unsurprisingly perhaps, the desirability for indexers to possess a wide general knowledge, and for indexers working within a specialist field to possess a specialist knowledge of that field.

**Skimming text**

The model of text comprehension and reduction so far described presupposes that reading takes place at the normal reading rate of 200–300 words per minute, which back-of-book indexers may or may not be able to adopt, depending on the proximity of deadlines, but which is economically impossible for academic indexing. Research into reading comprehension has shown that skilled readers can achieve overall comprehension at speeds of up to 600 words per minute, but with some loss of memory for details. At higher speeds text is skimmed rather than read, and experiments, confirming the accounts of indexers¹⁰, have demonstrated that skimming involves scanning text for perceptual cues. On a perceptual level, it is known that (1) a reader’s gaze duration increases linearly according to the number of letters in a word—i.e., long words are more readily noticed than short words; (2) uncommon words, and words with an unusual pattern of ascenders and descenders (as ‘yacht’) tend to be noticed; (3) deliberate visual effects, such as illustrations, tables, headings, italicized words, paragraph indentation, are noticed; and (4) definitions appearing in a text tend to be noticed, and may be over-emphasized by novice readers. More than one strategy may be applied simultaneously, and it has been argued that a skilled indexer processes at least three types of cue simultaneously in order to pick up definitions when scanning text¹¹.

On a conceptual level, skilled readers skimming text apply a number of selective processing strategies. (1) They may apply their knowledge of text structures to the seeking out of goal-significant areas of text. (2) In respect of individual words, semantic priming not only makes frequency of occurrence a significant cue, but also ensures that, once a word has been primed, semantically related words are also more likely to be
noticed. (3) If an important statement is sampled, the reader may choose to read it carefully. (4) 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. (5) Finally, a reader may draw plausible inferences to help connect propositions. Van Dijk and Kintsch describe these techniques as macrostrategies, which are more process-orientated than their macrorules: they 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.\(^{12}\)

**Index production**

Kintsch and van Dijk include text production in their model. In their view, the simplest operation in text production is the reproduction of particular words and phrases; and while most indexing consists of much more than this, ‘simplistic’ natural language indexing is essentially just that. Whatever the style of indexing, in cognitive terms, a set of index entries is among the most primitive and straightforward of all types of text production. There are two principal reasons for this. (1) In all normal indexing situations, the production of index entries follows immediately on from the comprehension of the text, so that loss of information through interference or decay is minimised; and furthermore the level of comprehension required is directed solely towards that task. (2) Compared with normal discourse, index terms are largely devoid of syntax, enabling the indexer to concentrate on the semantics of index entry production.

There are, however, other factors in modelling text production, two of which generate interference of one kind or another. One factor is the possibility of introducing metastatements. These are personal glosses, which may (but need not) be subconscious, expressing opinions or attitudes. They are most easily found in back-of-book indexes, and brief notices of instances are frequently found as fill-in notes in *The Indexer*. More seriously, Bell\(^{13}\) has described the problems indexers face in avoiding metastatements: “Indexers must ensure that the attitudes implied in the index accord with those of the text. We must choose to use ‘Terrorists’ or ‘Freedom fighters’, ‘Crowd’ or ‘Mob’, ‘Street riot’ or ‘Protest’, ‘Refugees’ or ‘Illegal immigrants’.’ The other factor generating interference is the erroneous reconstruction by a misapplication of world-knowledge of information that has been deleted from the indexer’s macrostructures—meaning, in effect, that the indexer gets hold of the wrong end of the stick. Thus a pamphlet entitled The *conservation of limestone pavements* was once classified with works on highway engineering by a British Library classifier who had clearly never experienced the pleasure of walking the Yorkshire Dales.

As well as general world-knowledge and specialist subject knowledge, three further kinds of knowledge are significant in index production. The first of these is knowledge of text structures. Research into text linguistics has described a limited number of text patterns (schemata), which many readers will recognize intuitively. A simple pattern is Problem—Solution, which in scientific papers is often expanded into the familiar pattern of Purpose—Method—Results—Conclusions. Another variation on the Problem—Solution pattern is one known as Multilayering\(^{14}\), in which one or more unsuccessful attempts at solving the problem are described before the text concludes with a successful resolution of the problem—a pattern illustrated on a very simple level by the story of the Three Little Pigs. Another approach to text structure is based on the observation that many texts are constructed hierarchically—taking us back to macrostructures—so that thematic sentences tend to appear at the beginning of paragraphs or sections of text. Hierarchical construction is seen at its most explicit in the construction of reports.

Next, indexers need to know the indexing systems they are using. For back-of-book indexing this includes any specifications or limitations imposed by the publisher, together with the indexer’s normal professional considerations of entry structure, filing arrangement and so on. The systems used in academic indexing vary greatly in their complexity. All controlled language systems by definition impose some controls on index terms, and even free language systems may well be subject to managerial constraints on exhaustivity. Indexers using controlled language systems will at the very least be constrained by the system’s requirements in respect of generic level and synonym and word form control. Precordinate indexes are the most complex, as their syntax is necessarily governed by rules. Experienced indexers carry a large part of their system’s controlled vocabulary in their heads, making text reduction and index entry production virtually one and the same process.

Finally—a much discussed aspect—the indexing process needs to take into account the requirements of users of the index. Indexing is based on prediction of the users’ requests, as Crystal\(^{15}\) has recently reiterated; it can be instructive to see how the same research
paper can be indexed quite differently in databases in different disciplines—for example in ERIC and PsycInfo. Problems really begin when a database tries to be all things to all men—and women, and children; the British Library's PRECIS system used to assign the subject heading MARINE ECOSYSTEMS to children's books on life under the sea.

Conclusion

The account of the indexing process outlined here is largely derived from research in psychology and linguistics, and the relevance of this research to indexing is based on analogy and anecdote. Todd's\textsuperscript{16} caveat that the macrorules have not been tested by indexers is a valid one, though the theory does appear to be compatible with indexers' own accounts of their cognitive processes. (A further instance has been supplied by Hazel Bell\textsuperscript{17}, whose criteria for selecting items for a main index entry, a subhead or a bare page reference can be readily interpreted in terms of deletion rules and macrostructural hierarchies.) A more general caveat is that a number of objections have been raised against the theory \textit{per se}, in particular that it does not explain the working of the macrorules that either transfer a proposition to the macrostructure or else jettison it. Others have cast doubt on all theories of discourse analysis and comprehension, on the grounds that none of them can be disproved. For all that, van Dijk and Kintsch's theories have an immediate attraction to indexers: indexing is text reduction, and we have here the fullest theoretical explanation yet of this process. While the macrorules have yet to be tested by indexers, they have provided the background for some recent experimental work by Endres-Niggemeyer\textsuperscript{18} on the process of abstracting. Her strategy, which could be equally applicable to indexing, was protocol analysis: the researcher gets the subjects to talk through their processes and actions as they are performing them. Another possible strategy would be the detailed examination of back-of-book indexes in order to discover if they are based on predicted patterns of text structure. Until some basic research is carried out, our knowledge of the indexing process must remain speculative.

References

3. There is a fuller account with additional references in my article 'Indexing as a cognitive process' in \textit{Encyclopedia of library and information science}, vol. 53, supp. 16. New York: Marcel Dekker, 1994, 155-71. Parts of the \textit{Encyclopedia} article are reproduced here with the permission of the publisher, Marcel Dekker Inc.

John Farrow is Senior Lecturer. Department of Library and Information Studies. Manchester Metropolitan University.

Facts on File

We have had some experience over the years with various publishers, both as authors and indexers. Rather than constantly berate publishers, we would like to present a bouquet to Facts on File. This New York publishing house has its own Indexing Services section, controlled by a Managing Editor, and produces a 16-page (plus index!) booklet on indexing. It expects very high standards of its indexes, but (and here we go again) they have a rule that authors must pay for the full costs of indexing or produce an index of the required standard themselves. This does at least help to maintain reasonable indexing standards, and it has to be said that authors' advance royalties are (slightly) more generous than from most British publishers of archaeology.

Lesley and Roy Adkins