REFLECTIONS ON ACTIVE AND PASSIVE VOICES IN INDEXING

David Batty*

ABSTRACT

Examination of index languages and systems of the last hundred years confirms that nearly all are based on an entity/activity combination. This combination is based on the use of phrases that reflect the passive rather than the active voice in natural language. Linguistic studies reveal a difference in the way the mind processes active and passive sentences and lends theoretical support to many pragmatic solutions to indexing problems.

Of the activities that have characterised our profession in the one hundred and more years of its formal existence, one of the earliest and, ever since, one of the most carefully considered, was and has been indexing. It has taken many forms, from simple natural language to the intellectual complexities of faceted classification and the statistical complexities of automatic indexing, and it has appeared in many situations, from the simple card index to the interactive computer system. During that same period the patterns of growth and the structure of knowledge have changed; a hundred years ago it was possible to draw a map of the universe of knowledge composed of hierarchies of discrete disciplines, but now as often as not we synthesise new multi- and interdisciplinary areas deliberately to meet specific demands — our own discipline of information science is a fairly

* David Batty, President, College of Library and Information Services, University of Maryland.
typical example.

In the development of index languages, from Dewey’s Decimal Classification and Cutter’s Rules for a Dictionary Catalog, to Thesaurusfacet and PRECIS, there has been a recognition of that increasing complexity of knowledge, and therefore of recorded knowledge. A corresponding attention has been paid by librarians and information specialists to the expression of complex index statements, and, especially in pre-coordinate indexing systems, to the order of the components of those statements.

This paper is concerned with the principles of order worked out and mostly held in common by the devisers of indexing systems, from the earliest time to the present day. The theme, it may be stated at the outset, is that in the light of our present understanding of theory, we may discern the same fundamental principles in early indexing systems as now. It should be understood that “indexing” includes all types and forms, alphabetic or systematic, manual or mechanized, though the arguments in this paper obviously apply less to any system in which single, independent terms, associated by common reference to a document, are not assembled in ordered strings – in those cases the arguments bear on the problems of search statements and search strategies.

The superficial problem of the order of a subject and its location in time or space, or the form of the document discussing it, were acknowledged and dealt with in almost tacit agreement by all devisers of index languages. Few went to any great length to say why subject should precede place, except to observe, correctly of course, that in most information enquires the subject is of primary importance, and less frequently is place paramount – in those cases when place refers to an area now made real in the anthropocentric context of knowledge as being man’s environment. For example in the Library of Congress List of Subject Headings places may be the main part of the heading when a document deals with the functioning of a region or city, and
sample headings like CHICAGO are fitted with a set of subdivisions by function.

The more important underlying problem of the order of the entities and activities intrinsic to a complex subject was dealt with almost implicitly by Dewey and explicitly and at some length by Cutter. In the very first edition of the Decimal Classification Dewey instructed that each language in class 400 PHILOLOGY should further be subdivided by the problems and process as listed in the first subclass 410 COMPARATIVE: an entity/problem combination. In other classes he instructed entity/problem combinations to be classed with the entity, as in 630 AGRICULTURE so that diseases of specific plants were classed at the plant without explicit subdivision. Cutter entertained similar arguments, but he was dealing with natural language, and the existence of the gerund and the customary formation of noun/gerund phrases like BULL FIGHTING or SERIALS CATALOGUING prompted him to favour natural language as a basis for his subject headings. But while his concern with the structure of natural language was fundamentally correct, it was unfortunate that he dealt mostly with the appearance and too little with the meaning of the structure.

The frequency of phrases like COMPUTER CATALOGUING and the similarity of the noun/participle and noun/gerund combination could only strengthen these assumptions. Further, in adjective/noun or adjectival-noun/noun phrases in English, Cutter saw in natural language an expression of his alphabetico-specific principle, since the adjective modifies the broader concept represented by the noun to reduce it to the more specific concept, as in CHILD MEDICINE. Admittedly, phrases like BLACK SHEEP or SCOTTISH FOLK SONGS could not be treated in this direct way, but it could be argued that the first example begins with an insignificant word, and would be represented simply as SHEEP, and the second example is already dealt with by the rule about subjects and places.

Around the turn of the century there was considerable
interest in the systematic ordering of index statements. Beginning in 1896 Paul Otlet made explicit and dynamic what had been in Dewey's *Decimal Classification* a convenient mnemonic device and space saver. His scheme, later the *Universal Decimal Classification*, proposed the synthesis of any subject or subdivision, one with another. It should be remarked here that the complexities of UDC are concerned mainly with auxiliaries like place, time and form and less with the internal structure of subjects, though the relational devises of UDC can be used to link entities and activities.

In 1906, J. D. Brown offered a radical classification scheme, the Subject Classification, whose fundamental principle lay so far ahead of his time that even he was inhibited by the professional climate of thought of his time from realizing its full potential. Brown rejected a basic organization by discipline, in which ROSES would be scattered among BOTANY, HORTICULTURE, etc., and instead proposed organization by the entity or "concrete," so that all works on ROSES would be gathered at a single location, subdivided by processes or problems, and even by aspects that in other schemes were the disciplines — the BOTANICAL ASPECT, etc.

The main classification schedules contained the major disciplines for the purpose of organizing the library shelves, but specific entities were listed only in the class of their first appearance in the scheme. Other classes where they might appear in other schemes contained only the activities proper to those classes. Many of these activities were repeated in a single, subsidiary table, to be used as subdivisions with the entity; in case they were not so repeated, Brown arranged a simple method of combining any two pieces of notation, to bring an entity and an activity together.

J. O. Kaiser proposed a method of systematic natural language indexing that ignored the natural language order favoured by Cutter, and instead accepted a formula of concrete/process/place that did something at least of what Cutter had tried to do.
Kaiser's method is simple but effective especially in the context of the card indexes he described, so that PAINTING SHIPS must always be expressed as SHIPS, PAINTING, with a left hand guide card for SHIPS and a centre guide card for PAINTING along with all the other activities concerning ships. Kaiser's method shares the fault with Cutter's that the same order is used for noun/gerund and noun/participle combinations that are similar in appearance but different in meaning — but at least the clarity of Kaiser's formula removes any doubt about the order of assembly if there is a conflict between the natural language order and a logical order.

At the heart of the entity/activity combination that we have so far observed as common to most if not all early index languages — and indeed the later ones as well — is the recognition of the entity as passive and the activity as something performed on that entity. This was implied in earlier remarks about noun/gerund and noun/participle phrases. In the phrase COMPUTER CATALOGUING, COMPUTER is not the entity being catalogued, and properly we should organize the phrase (regardless of how we finally express it) as [SOMETHING] CATALOGUING [BY] COMPUTER, since COMPUTER is an active agent. The same principle may be applied when two activities are involved in a complex statement; one of the activities becomes a passive system acted on by the other, as in ANALYSING CATALOGUING — better expressed as CATALOGUING, ANALYSIS. Notice here the translation of the gerund into a noun for the sake of clarity. Austin recommends such a practice in PRECIS indexing, and it should be remembered that the -ING forms were a principal element in nineteenth century alphabetical indexing, and a principal cause of ambiguity.

What we see here is the reflection in indexing of the passive voice in natural language. Nor is it surprising, especially when in indexing we must analyze and express complex topics in clear and systematic ways. Bar-Hillel refers to an insight of Victor Yngve when he says:
... by changing its mood from the active to the passive, the syntactic complexity of a given sentence may be reduced. And I have no objection to formulating this insight in the form that there exists a passive in English ... in order to allow ... the formulation of certain thoughts in sentences of a lower order of complexity.

In other words, active sentences, in English at least, are more intense and immediate, in which the function of the components may be understood only in relation to the sense of the whole sentence. Fenollosa made a similar suggestion in his examination of the relationships of word order and meaning in English and Chinese, and it may be observed that inflected languages whose formal word orders do not follow the sequence of described events relax their rules more in spoken forms must be understood as they are uttered.

Behind this realization lies the linguistic research of the last fifteen or twenty years. About 1960 Victor Yngve developed a "depth hypothesis" in which he distinguished among left-recursive sentences (which we would liken to the active voice), right-recursive sentences (the passive voice) and self-embedding sentences (a mixture). For example, MARY'S FRIEND'S SISTER'S CAT is left-recursive, because the chain of possessive nouns leading to CAT precede it, while THE CAT OF THE SISTER OF THE FRIEND OF MARY is right-recursive, because the chain of possessive nouns lie to the right of CAT. THE CAT THE GIRL SAW IS OUTSIDE is a self-embedding sentence; the sentence THE GIRL SAW THE CAT is now embedded in THE CAT IS OUTSIDE. Yngve's hypothesis was that left recursive or active structures are more difficult to process than right-recursive because all of the sentence must be scanned by the mind to determine the function of any element, whereas right-recursive sentences may be scanned and understood sequentially. If we describe the mental processing of sentences in terms of automated theory, then we can distinguish between a short term memory of very limited capacity used mostly as a processing area, and a long
term memory used as mass storage for the rules of syntax, and available as an overflow or an overtaxed short term memory. Actives sentences may become too much for the short term memory to handle as a single unit, but passive sentences may be broken down to use long term as well as short term memory. Chomsky has objected to this distinction, on the grounds that self-embedding sentences cause the greatest difficulty, because the mind is required to hold one sentence while analyzing another within it, and Chomsky and George Miller have further suggested that the mind cannot easily perform an operation if it is already performing that same operation on another sentence. However, these objections do not necessarily imply that the mind processes active and passive sentences in the same way, and indeed, Lyons, in his book on Chomsky, describes experiments that demonstrate that active sentences are more readily understood as a whole than are passive sentences. In the PRECIS manual Austin refers to Lyons' assertion to support the PRECIS format, in which the lead term and the qualifier represent an active phrase, and the analyzed PRECIS string remaining in the display represents a term-by-term layout in the passive form.

The realization of the difference between active and passive voices reveals the source of much of our instinctive indexing, and indeed of much inventory descriptions. It also emphasizes the care with which we must examine the phrases we use in indexing. BIRD MIGRATIONS appealed to CUTTER because of its natural language order; it also appears to satisfy our requirements for an entity/activity order. On closer inspection we may wish to object to it, because it is the birds that migrate, and we may wish to substitute a passive phrase for what now appears to be an active one. But on further reflection we should say that MIGRATION is not caused by BIRDS: properly speaking, BIRDS are the passive entity and MIGRATION is what happens to them. On the other hand, STEAM PLOUGHING is like the earlier example of COMPUTER CATALOGUING, since it really represents [LAND] PLOUGHING [BY] STEAM [PLOUGHS]. To aid such an analy-
sis of language, Austin has turned to deep case grammar to enable the representation of the proper relationships of terms in the PRECIS string. In this way he is seeking to avoid these confusions raised for present day indexing by reliance on surface-structure.

Adoption of the entity/activity combination and of the passive rather than the active phrase was demonstrated empirically at least to be fundamental to effective indexing, and linguistic theory may offer further theoretical justification.

The first part of this paper spent some time on early systems of indexing and classification; that concentration was deliberate. It is ironic to note that by 1910 all the types of pre-coordinate indexing later developed to highly sophisticated levels had already been given practical expression, and all of these included those fundamental principles, albeit sometimes in primitive form. More ironically still, these very powerful theoretical bases all had very pragmatic, even physical origins: Dewey's relative shelf location to replace fixed location, Cutter's recognition of the needs of libraries of different physical size, Brown's support on open access and browsing, Kaiser's need to organize card index file drawers, Otlet's concern to handle files of ephemeral material. Later, highly complex schemes of classification and indexing may be seen as ever more elaborate forms of these types, culminating, in classification at least, with the intricate structures of Ranganathan's Colon Classification - the end of a tradition rather than the beginning of a new one. In the present decade, with the application of linguistics to index language development and operation, and with the manifestation of new and clearer syntheses of indexing principles in systems like Thesaurofacet and PRECIS, we may well say with T. S. Eliot:

And the end of all our exploring
Will be to arrive where we started
And see the place for the first time.
Bibliography


Brown, J.D. *Subject Classification.* London: Grafton, 1906.


Ellot, T. S. *Four quartets.* London: Faber and Faber, 1944.


