III S Y N T A X

III . 1 Input syntax

Syntactic information in document descriptions may be implicit or explicit. In treating keyword strings or compound subjects as units, we rely in matching on the presence of relations between the elements involved, though these are not defined. Alternatively we may make the relations explicit, minimally by ordering or bracketing, more fully by specifying particular relationships. Analysis may accordingly be intended simply to identify word strings which function as units, or to spell out the relational structure of the input text. In the first case partial parsing procedures may be adequate; in the second full analysis may be required.

Syntactic analysis may also be exploited to identify text items with specific syntactic functions, on the grounds that these are key content indicators.

Most automatic syntax analysis in document retrieval comes under the heading of partial parsing. Few projects have attempted full parsing. This is more usual in question answering systems.

III.1.1 Partial parsing

The lowest level of partial parsing is represented by the use of specific, very restricted clues as aids to the selection of particular items. Borkowski's 1970, 1973 work on case citations in legal texts, in which "v." is used as a trigger to pick up related items, is an example. A similar approach using templates of the form 'from ... to ...' to identify dates is suggested by Weiss 1969.

The next level is characterised by the use of a wider range of clues to process complete sentences rather than merely to select certain items. For example we may use punctuation marks, prepositions and conjunctions to divide sentences into strings of words which may be treated as units. Titles naturally lend themselves to this treatment, and it has been attempted in Armitage's 1967, 1968, 1970 work on articulated subject indexes. This will be discussed later.

In this case there is no attempt to characterise the phrase structure of the separate strings identified. Clarke and Wall's early project, 1965, went a step further in parsing to identify phrases, and specifically noun phrases, though they did not try to relate the phrases to one another to produce an overall sentence characterisation. Carroll 1973 reports without details the use of a text analyser to identify noun phrases for profiles. Since the objective is limited, the analysis is presumably fairly simple.

The procedures described by Klingbiel 1973a,b are intended to be applicable to the vast quantities of material received by the DDC. They have already been tested on millions of words of abstract text. The syntax analysis routines select words or word strings for possible use as index descriptors. Text is segmented, for example at punctuation marks. Individual words are checked against a dictionary, and sequences of syntactic categories for selected words are vetted by reference to a format dictionary of 76 entries. Word strings passing the format test are offered as candidate index terms for human review. The examples of output given in Klingbiel 1973b are quite plausible, and the scale of the whole operation is very impressive. However retrieval procedures have not yet been developed.

The next level is reached where a sentence is substantially or completely parsed, but the overall refinement of the structural description is not very great because only limited information is to be selected for use in document description. Approaches of this sort verge on full parsing.

One project of this type is Hillman's 1968, 1969, 1973. It is difficult to be certain of the details, since there seem to have been some changes in the techniques used. Hillman's approach is fairly sophisticated: it is designed to identify noun phrases, which are assumed to be the main content bearing text items, and also their relationships. A relatively simple phrase structure parsing of a sentence is therefore processed to discover the logical relations holding between noun phrase units, chiefly by looking at verb environments. The analysis output is a set of canonical components representing propositions expressing relations linking noun phrases, to which the original sentence has been reduced. This relational information is not, however, preserved in document descriptions, but is merely used to generate weights. Though the analysis is more complex, therefore, Hillman's overall objective is the same as that of the other projects mentioned, namely to identify semantic units. Hillman 1969 reports work on 1000 document texts, and Hillman 1973 describes an operational retrieval service with substantial date files, but there is unfortunately no indication either of the accuracy of the analysis procedure or of the value of its output.

The Syntol project (Syntol 1964, 1965, 1967, 1970) differs in that the relational information extracted in analysis is carried over to indexing. It has made the most consistent and ambitious attack on syntactic analysis and indexing, designed to generate document descriptions similar to Farradane's, and the attempt to produce such sophisticated descriptions automatically is clearly of considerable interest.

The Syntol research is described in some detail in Sparck Jones 1973a. It is based on the assumption that the syntactic analysis procedures required to identify the logical relationships to be indicated in document descriptions must be based on relatively detailed rules and dictionary information. Initial attempts to work with a rather crude analysis procedure using clue words, described in Syntol 1967, were not discriminating enough. However the syntax rules used subsequently were not derived from any systematic view of grammar, but were developed ad hoc for French, and may well have been influenced by the particular subject field, psycho-physiology, in which the system

has been tested. Though the analysis procedure is not required to provide a full structural picture of sentences, it has to be carried far enough to pick up key concepts and their relations, and this implies fairly detailed processing.

The experiments reported in Syntol 1970 were carried out on 1016 abstracts. They involved a dictionary of about 7000 words or word groups with detailed morphological and syntactic information, the latter invoking 57 syntactic categories, for example. The initial function of the syntactic analysis routines is to identify word groups and to resolve ambiguities open to syntactic treatment. (There is concurrent semantic processing to select lexical items by reference to a descriptor dictionary of some 3000 items.) The tests showed a generally satisfactory recognition of groups, and over 90% success in ambiguity resolution. The second stage of syntactic analysis is designed to identify the syntactic function of specific words or groups, and the type and structure of the phrases. For example, the input sentence "Modification du comportement du chien après ablation du lobes orbitaux et frontaux" is divided into an initial subject phrase, three prepositional phrases, and a final coordinate phrase. Some 750 rules for exploiting the appropriate dictionary information are required to carry out the analysis. Inspection of a sample of 155 abstracts showed few absolute failures of analysis, though some phrases received several interpretations. However the success of the procedure is really to be judged by the descriptions generated, to be considered below.

III.1.2 Statistical syntax

This can be described as back-door partial parsing. There is no attempt to characterise syntactic structures occurring in the input text in the usual way. The assumption is simply that if two content words tend to cooccur within a restricted context, like a sentence, this must reflect the presence of a syntactic or logical relation between them. Pairs or n-tuples of words thus cooccurring may therefore be selected to function as complex descriptors. The main difficulty is picking up significant frequencies of cooccurrence, since it is quite a strong requirement that two words should cooccur within a sentence or other frame sufficiently frequently within a text for their cooccurrence to be significant.

Early work on the Smart project (Salton 1968a,c) attempted to identify "statistical phrases" intended for use as complex index terms. The process was not genuinely statistical, since reference was made to a dictionary of acceptable word, or rather thesaurus class, combinations, cooccurrence of members of the relevant classes in a sentence being the criterion for the assignment of the corresponding descriptor. Retrieval tests with three small collections showed the procedure was not especially helpful, performance being the same as that of a simple thesaurus. Salton attributes this to the restrictive prior dictionary rather than false combinations.

A rather similar approach was tried by Artandi 1969a,b, for the automatic generation of links. It was restricted to relating words of two types, drug names and 'modifiers' like "effect", defined by an indexing vocabulary. Tests with 15 document texts showed 63 out of 285 incorrect links. These were generally associated with greater inter-word distances, but Artandi concludes that simple distance criteria would probably be too crude for full texts, where style varies, though they might be applicable to abstracts. These experiments were on too small a scale to be particularly informative.

III.1.3 Full parsing

Full parsing for document description appears to have been seriously attempted only in the early stages of the Smart project. The procedure is fully described in Salton 1968a. Its object was to identify phrases, like "retrieval of information", which could lead to the assignment of complex descriptors. The Harvard Predictive Analyser (Kuno 1965, 1966) was used to give a phrase structure parsing of sentences. Specific substructures, for example those representing a subject-verb relationship, were then matched against a dictionary of 'criterion phrases'. A criterion phrase represents an acceptable combination of words, or thesaurus classes, with syntactic dependency relations between them. Each phrase characterises a type of structure which may be instantiated in a variety of specific ways in text. For example a single noun+qualifier criterion phrase with the qualifier dependent on the noun, like 'house+brick' could occur in actual text as either "brick house" or "house of brick". The dictionary comparison is not trivial, since allowance has to be made for text and dictionary structures which match in essentials but not in every detail.

At the time, experiments with the Analyser were relatively expensive and only a limited number were conducted. The results of actual retrieval experiments exploiting the technique were also not very promising: with one small collection it turned out that these syntactic phrases performed less well than the statistical phrases described earlier (and hence than a simple thesaurus). Salton's view is that the inadequacy of the grammar is probably responsible, and reliance on a prior dictionary may also account for some failures.

III.1.4 Conclusion on input syntax analysis

Three points should be made. The first is that even the most ambitious approach, the Syntol one, is orientated to some selection of information from the input text. In question answering, in contrast, all the information in the input text may be extracted for future use in retrieval. The second is that only a few of the projects are designed to extract relational information which will actually be explicitly indicated in document descriptions (as opposed to implicitly in compound descriptors): the Syntol workers and Artandi represent two extremes here. The third is that all parsers used are,

by linguist's standards, inferior, since they are all of the phrase structure type. Special problems from the documentalists' point of view are presented by analysis procedures like the Syntol one, or Salton's, which require an elaborate word dictionary. The effort involved in constructing such a dictionary is substantial. The idea of distinguishing closed and open word classes, and of using a limited dictionary confined to closed class words like prepositions and conjunctions, and perhaps suffixes, is an attractive one. The hypothesis is that information about these words can be exploited when they occur in sentences to assign syntactic categories to other words, and hence permit parsing. The practical advantages of reducing the parsing dictionary in this way are clear. The idea has been exploited for full parsing, for example by Thorne 1968; it appears to be particularly well suited to documentation where only partial parsing may be required. In fact Hillman 1968 adopts this approach, and it is also followed by Earl and Rush in their work on extracting.

In general, it will be evident that it is extremely difficult to comment on the effectiveness of the analysis procedures described: evaluation really depends on the value of either the selected descriptors or the identified relationships for retrieval.

III . 2 Description syntax

As noted, the objective of input syntactic analysis is often simply to identify key items for use as descriptors in indexing. There is no intention of providing syntactically structured descriptions except implicitly, where complex descriptors are used. Salton treats his phrases in this way, for example. In general in indexing the treatment of compound descriptors, or precoordinate subject specifications, may be more or less refined: a string of words or terms may be regarded simply as an indissoluble whole, or some distinction may be made between main and subordinate elements. The latter, to be found in printed indexes, is of course a move towards explicit syntax. It is difficult to get much idea of the complexity of strings extracted automatically: Salton 1968a mentions phrases like "computer control"; an example in Hillman 1973 includes both "rock bolts" and "roof support in underground excavations", and the output given in Klingbiel 1973b contains "interactive retrieval" and "automated indexing techniques".

The main project which has sought to express syntactic functions and relations explicitly in document descriptions is the Syntol one (1964, 1965, 1967, 1970). A Syntol description consists of one or more 'syntagms': these are pairs of index terms linked by general relations like cause or association. The Syntol workers do not advocate any particular relations: a Syntol-type language might use anything from two to twenty broad logical relations. The tests reported in Syntol 1970 used three, the consecutive, comparative and associative relations. The last is the least restrictive and is used as a sort of residual relation if neither of the first two holds. The detailed procedures for creating syntagms for the first two relations thus differ from those for the third.

The second major component of the Syntol document processing package is therefore the procedure for translating semantic and syntactic information associated with the phrases identified by the analyser into syntagms. As indicated earlier, individual words are replaced by index terms taken from a substantial hierarchically structured thesaurus. Extraction of the first two relations involves the use of clue words, like prepositions and verbs of certain sorts, and syntactic schemas specifying characteristic constructions associated with the clues. If a clue and schema match an analysed syntactic structure, the appropriate logical relation between indicated terms is established. Semantic checking involving a 'réseau notionnel' which characterises collocational relations holding between thesaurus descriptors may be required. Associative relations are normally supplied if consecution and comparison are not established. The apparatus required is fairly complicated: for example there are 30 schemas for the type of transitive verb clue represented by "provoquer".

The output from the procedure can be illustrated by an example. The abstract text "Chez dix singes stimulés par électrochoc cortical direct et chez un malade la procaine intraveineuse a protégé de l'épilepsie pendant une demi-heure; la xylocaine a un effet moindre" gave syntagms depending on associative (A) and consecutive (C) relations as follows: PROCAINE C CRISE, PROCAINE C EPILEPSIE, PROCAINE A PROTECTION, PROTECTION A EPILIPSIE, DIRECT A CORTEX, ELECTROCHOC A STIMULATION, STIMULATION A ETRE, XYLOCAINE A - . The test results obtained were evaluated by comparison with manual indexing for a sample of the abstracts processed. The syntactic procedures for generating syntagms gave 97% correct results; but, surprisingly, the semantic checking reduced this to 73%. Some details of the evaluation are unclear; and it is worth noticing that nearly 20% of the indexing expressions consisted of isolated thesaurus descriptors, and that 80% of the syntagms involved the very general associative relation.

At a lower level, Artandi 1969a,b generates index descriptions involving links between controlled language terms. As noted, 22% of the links found in 15 document texts were judged incorrect.

III . 3 Index language syntax

Recent general discussions of index language syntax are to be found in Gardin 1973 and Coates 1973. Syntax appears in an indexing language a) in the indexing vocabulary if this consists of compound terms or subject headings, or b) in the set of relations which may be used to link terms. For manual indexing there is a large literature on the construction of subject vocabularies, and on techniques for explicitly indicating relationships, for example by the UDC ':', facet order, relational operators (Farradane 1967, 1973), etc. I shall ignore this here.

Fully automatic text derived syntactic indexing has not been much attempted. Hillman's 1968, 1969, 1973 approach is derivative, but most of the approaches mentioned rely on the control imposed by a dictionary, either indirectly through word or term selection, or directly as in Salton's 1968a criterion phrase technique.

III . 4 Search syntax

Machine searching of manually assigned subject headings is not in question here.

With compound descriptors, any procedures which select a component, say by exploiting a hierarchical relationship to replace XYZ by Z, might be claimed to be syntactic, but it is perhaps more appropriate to regard them as semantic.

When syntactic information is explicitly provided in document descriptions, obvious ways of modifying it during searching are by simplifying expressions and by weakening relations. For example a document characterised by the expression 'A rel B rel C' may be allowed to match a request consisting of 'A rel C'; alternatively, if we have a set of specific relations, rel_1, rel_2 and rel_3 with rel_3 weaker than the others, we may allow a match on rel_3 as a substitute for one on the others. In principle permitted operations should be defined by rule and not specifically by the provision of alternative forms of a request. Of course manually generated descriptions of this kind could be manipulated automatically in searching, as Farradane has suggested, but it does not seem to have been attempted, other than in an early stage of the Syntol work (Syntol 1964): the results of these experiments suggested that relations must be weakened to retrieve sufficient relevant documents. In the recent research on automatic Syntol indexing (Syntol 1970) there have been no actual retrieval experiments.

Clearly, if syntactic information is incorporated in document descriptions, the treatment of requests must correspond. Automatic request processing is illustrated by Hillman's LEADER system for interactive searching (1968, 1969, 1973) and by the Smart project (Salton 1968a). However if documents are indexed by straightforward keyword strings, these may be extracted manually from requests.

As noted earlier, Boolean requests can be said to have a syntactic structure which may not be explicit in document descriptions. Boolean requests are ordinarily prepared manually, and are very common in operational systems, for example UKCIS (Barker 1972a,b), IITRI (Williams 1972). A good deal of effort may be involved, and care, for instance in the treatment of not-logic (Scheffler 1972).

These remarks apply to the processing of submitted requests or profiles. Their original generation is a somewhat different matter. Carroll's 1973 experiments with parsing are of interest since they were designed to generate profiles from source documents. Retrieval comparisons between 15 machine and Boolean profiles for about 10000

documents showed improved recall with comparable precision for the machine profiles. The technique involved is rather like request modification by relevance feedback in iterative searching, to be considered later.

III . 5 Conclusion on syntactic indexing

The general value of syntactic information for retrieval has been strenuously argued. In the present context there are several distinct questions to examine.

We must first separate syntactic criteria from syntactic descriptions, that is procedures using syntactic information simply to identify semantically important items to be used for describing documents from procedures retaining syntactic information for its own sake. In this case the question is whether document keys thus identified are more valuable as descriptors than ones obtained without reference to syntax, for instance statistically. A subsidiary question is whether, assuming such keys are preferable, they can be adequately recognised automatically. In fact in manual indexing syntactic criteria may well be exploited, unconsciously, as a natural aid in reading text. But there is no obvious way of conducting a sensible experiment in the comparative automatic and manual use of syntactic criteria for selecting words. It is more useful to ask whether syntactically based automatic extraction methods are preferable to non-syntactic ones. Unfortunately there is rather little evidence to go on. Of the work described, such projects as Borkowski's, for which good performance is claimed, are too specialised to be relevant. Hillman does not provide any concrete performance figures for his system, or attempt to evaluate it other than sociologically, and Klingbiel has not yet attempted searching. Salton does provide some limited comparative results. As mentioned earlier, performance for his syntactic phrase procedure, over 17 requests and some hundreds of documents, is noticeably inferior to that of statistical phrases. Inspection of other performance figures in Salton 1968a suggests it is much like that of simple keyword stems taken from the texts.

In considering the value of syntactic information in descriptions, we can first ask whether the implicit syntactic information of precoordinate descriptors is of value, and secondly whether explicit information is helpful, with a further distinction between relatively sophisticated information like that provided by Syntol encodements and simple links. Again, if it can be shown that such information is of value, we can further ask whether it can be effectively picked up automatically.

It is not appropriate here to rehearse all the arguments for and against precoordinate descriptors, for the use of minimal syntactic links, or for more ambitious indexing of the type advocated by Farradane and implemented, for instance, in the Titus system (Bousselet 1973). The following points are perhaps sufficient. Evaluation experiments designed to compare, for manual indexing, precoordinate

subject indexing with post coordinate terms, or syntactically structured descriptions with simple term lists, suggest that the use of either implicit or explicit syntactic information in document descriptions is of no general material value. See, for example, Saracevic 1968, 1971, Cleverdon 1966, and Keen 1972, 1973. At the same time, postcoordinate subject indexes continue to be made, and systems like the Titus one exist. It must be allowed that syntactic descriptions may be of value in specific contexts, but there is no hard evidence to the effect that they are generally of value.

This being the case, the justification for automatic syntactic indexing appears dubious. The few evaluation experiments which have been conducted, like those of Salton's just mentioned, suggest that automatic syntactic indexing is neither better nor worse than its manual parallel. However, it may be that the correct way of providing syntactic information in document descriptions has not yet been discovered, and that when it has, it may be worth writing, or attempting to write, programs to provide it automatically.