THE

Journal of Documentation

VOLUME 23 NUMBER 4 DECEMBER 1967

AUTOMATIC DOCUMENT SELECTION WITHOUT INDEXING

W. J. HUTCHINS

Sheffield University Library

Given the demonstrable deficiencies of indexing and indexes as means of document analysis and selection, a system is proposed which matches uncondensed and unanalysed texts with search requests and semantically equivalent transformations derived from them. The method utilizes the results of machine translation and structural linguistics in syntactic analysis and in semantic classification with adaptations to the requirements of a document selection system.

1.1 AN INDEX is considered an essential tool for the full exploitation of a collection of documents once it has reached such a size that personal perusal of each text by every user for documents relevant to his needs has become laborious and impractical. The volume of literature on indexing systems, indexing theory, and automatic indexing is indicative of the importance workers in the field of information retrieval attach to indexing. Whether by man or machine the basic processes of indexing are: the analysis of each document's content, a formulation of this content in a set of descriptors, and an organization of descriptors such that enquirers can match their search requests and not miss any documents relevant to that request. Since enquirers express search requests in natural language, whether the content is primarily described in a code (e.g. in the conventional classified catalogue) or not, an index comprised of descriptors selected from the words of natural language must be provided and this index is the main guide to the collection's contents for the majority of users.

1.2 The inherent weakness in content-analysis has been well put by Yngve:[17] 'abstracting amounts to a careful manual searching of each document for answers to certain predetermined questions' and 'the system can give direct answers only to those questions that the abstracter has considered

and effectively answered. The system fails when the user comes with a question that has not been foreseen . . . The material left out can never be retrieved by the system.' In other words, each document is analysed and assigned descriptors according to criteria considered valid for documents already processed; these criteria may not be relevant now or in the future for certain enquirers. 'There is no absolute "Relevance" of a document. It depends on the person and his background, the work and the date.' (Mooers,[9] p. 2).

1.3 The deficiencies are well known of indexing systems which select descriptors solely from the texts of documents and do not relate one set of descriptors with another. Derivative indexing (which includes many automatic indexing systems) 'is constrained by a particular individual's personal manner of expression . . . is remarkably sensitive to a particular period of time . . . (and) the user has no advance knowledge of the terminology that has been used in all the various texts of a collection'. (Stevens,[13] p. 174.) Since the same subject or concept may be referred to by different words, the user will miss relevant documents if he does not know all the synonyms.

1.4 In order to facilitate the retrieval of relevant documents and to exclude irrelevant ones, indexing systems regulate their vocabulary and organization so that, as far as possible, descriptors stand in a one-to-one relationship with their denotata (things or concepts). When one denotatum is designated by more than one word (or compound word) as in synonymity, the index must select one of them. Its decision may be arbitrary since users of the language may disagree on the degree of similarity between the denotata of synonyms. In other cases, one word may designate different denotata (homonymity). Also users may differ in their attitudes to and knowledge of the denotatum of a word (polysemia)—thus while its designation is the same for members of the language community its connotations may be quite different, e.g. 'salt' has different connotations for a cook than for a chemist since their knowledge of the substance differs (Antal,[1] Morris[10]). In order to distinguish such diverging usages of the same word, index languages provide some kind of 'definition': often an indication of the context in which the word may be found, e.g. pitch (acoustics), pitch (angle), pitch (substance). In addition to the general problem of polysemia, indexers must decide the extent to which specialist vocabulary and word-usage is admitted in an index intended also for non-specialists. Specialists in many disciplines need to discriminate between the denotata of words which for other users of the language are near-synonyms. They also need to designate concepts or objects not recognized in the general vocabulary: thus they create new words (e.g. 'cybernetics') or new combinations of commonly known words (e.g. 'solid state physics' or 'nuclear magnetic resonance'). For the non-specialist both kinds of neologism may be meaningless; either the word (or this particular usage) is unknown or the combination of words may not designate for him any recognizable concept.

In his discussion of the problems associated with specialist vocabulary F. Jonker[5] concluded (p. 45) that 'since there is no universal viewpoint, there appears to be no basis for maintaining that universal index language is possible'.

1.5 The introduction of classification into an index—whether as a separate organization of descriptors (as in a classified catalogue) or as a system of cross-references—is necessitated by the restrictions placed on the number of descriptors permitted for each document. Without restrictions the index file would rival the document collection itself in size. Thus each descriptor designates a large class of denotata communicated in the document; it stands in a generic relationship to them. Another descriptor in the index may refer to some of these denotata but not to others (class intersection) and a third may refer to all and also to some others (class inclusion). Thus descriptors can be ranked in hierarchies according to the domain of their reference. If relevant documents are retrieved by a search request formulated in one descriptor, others may be retrieved by descriptors related to it in such a hierarchy. If no classification were provided, these other relevant documents would be missed. However, a wide variety of hierarchies of denotata are found in classification schemes; no agreement seems possible on a universally acceptable scheme, for the simple reason that there is no universal attitude to denotata (already demonstrated by the presence in language of polysemes).

1.6 Index languages rarely provide adequate syntactic organizations of descriptors. In some, e.g. co-ordinate indexing and KWIC indexing, descriptors are not linked and so no syntactic guidance is given. Users are expected to formulate search requests as 'bare' descriptors and to hope that if a document has been assigned matching descriptors the relationship between them will be the one desired; thus 'false drops' are likely. When descriptors are linked they are usually ordered in sequences, e.g. 'Concrete, Strength' and 'Books, Illustrated, Paper, Coating, Density'. The only relationship given is that of juxtaposition. Without the syntactic guidance of natural language (e.g. prepositions, articles, inflection) such descriptor sequences may be ambiguous (cf. Vickery's example:[18] 'Bacteria, Dyestuffs, Destruction') and no agreement on the optimum ordering seems possible. The introduction of non-linguistic guidance in the index is likely to baffle users, hence many indexing systems provide a separate syntagmatic organization. This is generally incorporated in a classified catalogue (as 'roles', 'links', 'facets', etc.) where the relationships between denotata can be expressed more explicitly and logically than in natural language (since the syntax of natural languages indicates relationships between word-signs and not between their denotata (Morris,[10] p. 27).) In Syntol[8] 'syntagmas' are provided outside the classification scheme, but still separate from the descriptors. In such systems users must formulate their request using 'logical' syntax rather than the syntax of natural language.

1.7 When we add 'the statistical fact that only a fraction of stored information is ever referred to again' and that consequently 'the intellectual effort invested initially in the interpretation of information for translating it into a common language (i.e. an index language) can only be partly recovered' (Luhn[8]) we are justified in looking for methods of document selection not involving the compilation or use of an index at any stage.

1.8 The requirements for such a system might be summarized as:

a the whole or a substantial portion (see paragraph 5.4) of the texts of documents in the author's own wording is to be searched.

b the search request is to be expressed in natural language.

c suitable transformations are to be made on the search request to find semantically equivalent formulations which may then be matched with expressions in the document texts. The rationale for this is that in order to decide whether a document treats the subject sought a reader scans the text for statements corresponding to his own formulation of the subject.

Thus the indexing processes are replaced by a system transforming the search request into equivalent expressions in natural language and matching them with phrases and sentences in document texts. A tentative system, called here for convenience QUANTRAS (an abbreviation of Question Analysis Transformation and Search), is put forward and then followed by a discussion of the major problems involved.

2 THE QUANTRAS RETRIEVAL SYSTEM

2.1 The basic components, stored permanently in the computer, are the following:

a *The Dictionary* which contains words of all languages accepted in the system (see paragraph 5.4). For each word a Concept number and a Word-class are provided. Compound words are treated as single words.

Concept numbers. Different numbers are given to distinguishable denotata

i All words differing only in their syntactic functions, e.g. the noun 'drama' and its (semantically equivalent) derivatives: the adjective 'dramatic', the adverb 'dramatically', the verb 'dramatize', etc. are assigned the same Concept number.

ii All synonymous words (see paragraph 5.1) are given the same Concept number.

Word-classes. Grammatical categories, roughly equivalent to the 'parts of speech' of traditional grammar but not necessarily the same. The detail in which the syntactic functions of words are analysed depends solely on the requirements of the automatic syntactic analysis program: thus while some parts of speech may be confounded others may be subdivided. The use of 'Adj', 'N', etc., in this paper is purely illustrative.

Where a word has more than one meaning (a homonym) it is given a Concept number for each meaning; where a word has more than one grammatical function, such as 'play' as a Noun or a Verb, it is assigned a Word-class for each function.

b *The Syntactic Analyser* which is a computer program, based on programs developed in machine translation, for the discovery of the phrase-structures underlying search requests.

c *The Syntax Pool* which is a listing of all the basic phrase-structures of the language, grouping them in sets of semantically equivalent phrases. One such set would include 'Adjective—Noun', 'Noun—*of*—Noun' and 'Genitive Noun—Noun' since these structures underlie such semantically equivalent phrases as 'French industry', 'industry of France', and 'France's industry'.

d *The Catalogue* which includes one entry for each document. As in a conventional library catalogue, entries may be bibliographically normalized and give little more than titles and subtitles. Or they may give (and for the effectiveness of the system this is assumed) more extensive extracts from texts such as abstracts, chapter headings, or even full texts (see paragraph 5.4).

2.2 Large storage is required for both the Dictionary and the Catalogue, which are consequently brought into operation only as necessary. Small storage capacity with high-speed access is needed for the Syntactic Analyser and the Syntax Pool and for other stores required temporarily during the processes.

The retrieval system is described and schematically illustrated in two stages: Stage I analyses the search request and provides for equivalent transformations, and Stage II matches these transformations against Catalogue entries. Each step is illustrated by a simple example.

2.3 STAGE I: ANALYSIS AND TRANSFORMATION

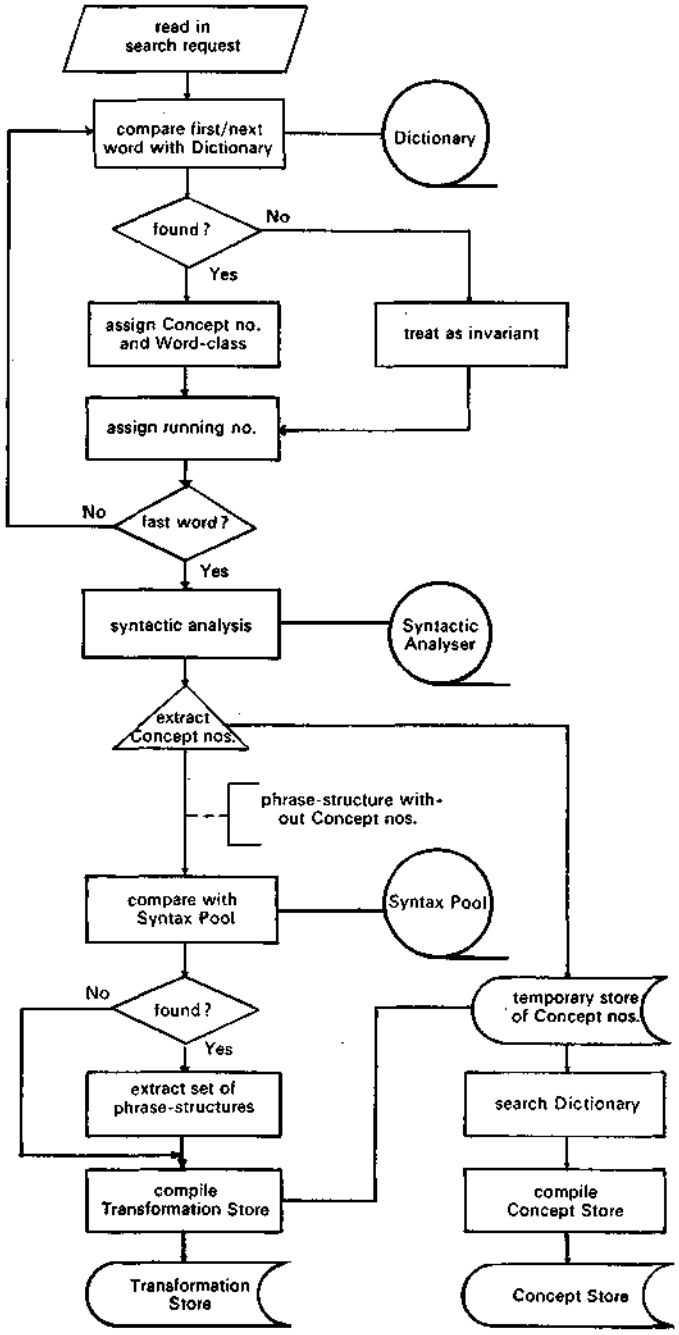
(1) The search request is formulated in natural language by the inquirer. Before it can be analysed it must be relieved of words not directly related to the 'subject content' of the request, e.g. such phrases as 'What has the library on . . . ?', 'Have you anything on . . . ?', etc. This process can be done by the librarian before input or possibly by procedures within the computer itself (e.g. comparison with a list of 'unimportant' words and phrases).

Example:

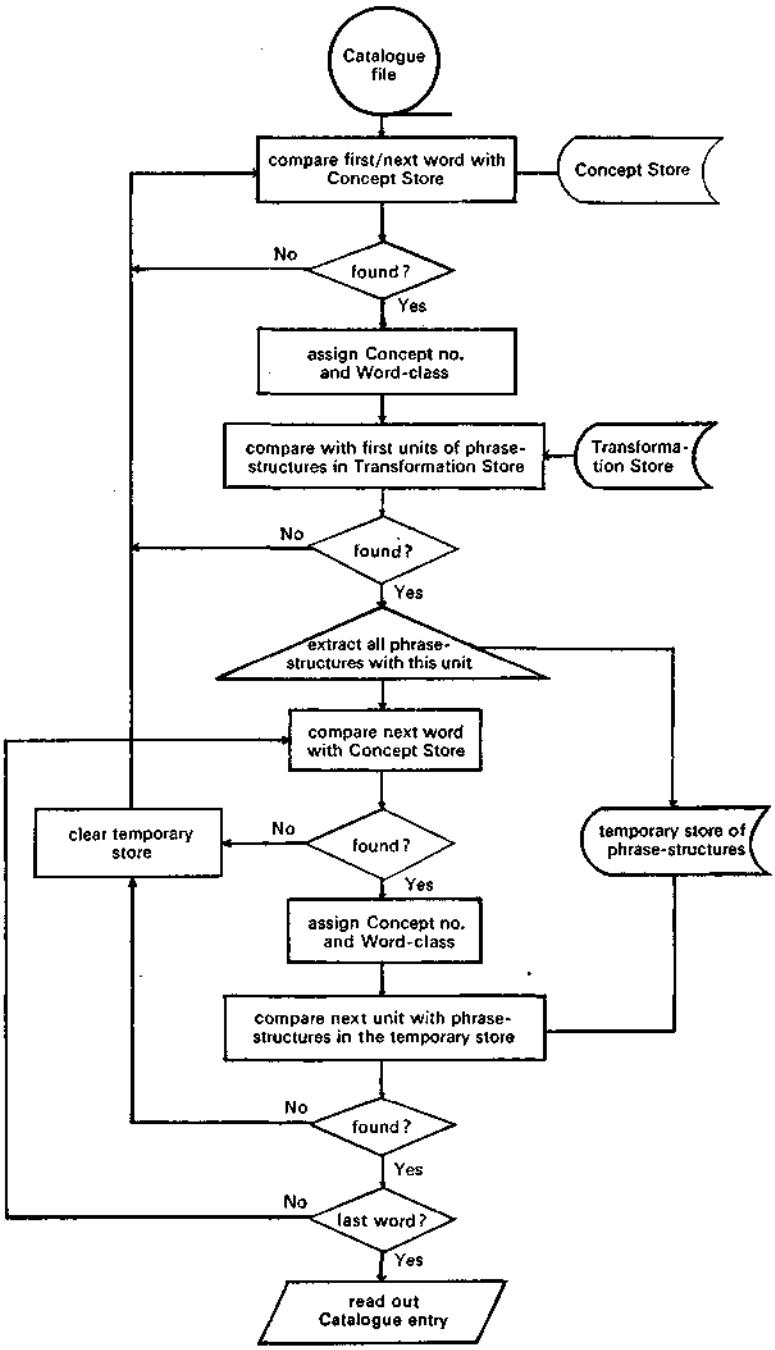
Is there anything about American economic history? → American economic history.

(2) Each word in turn is compared with entries in the Dictionary and assigned from it a Concept number and a Word-class (or alternative Concept numbers and Word-classes). Since 'compound words' are treated as units,

STAGE I. *Analysis and transformation*



STAGE II. *Search*



techniques devised for machine translation to deal with word-sequences (idioms as well as compounds) must be incorporated.

Example:

American economic history → $\begin{Bmatrix} 310,\text{Adj} \\ 310,\text{N} \end{Bmatrix}$ + 162,Adj + 253,N

(3) Each unit (Concept number/Word-class group) is given a running number.

Example:

$\begin{Bmatrix} 310,\text{Adj} \\ 310,\text{N} \end{Bmatrix}$ + 162,Adj + 253,N → $\begin{Bmatrix} 01,310,\text{Adj} \\ 01,310\text{N} \end{Bmatrix}$ + 02,162,Adj + 03,253,N

(4) On the basis of the assigned Word-classes, the phrase is analysed for its underlying syntactic structure. Only sequences of Word-classes which are derivable from the grammatical rules of the language will be accepted by the Syntactic Analyser. Thus a phrase with the sequence 'Adj+Adv+N+Ngen' (e.g. large beautifully house Peter's) violates the rules of English syntax. When alternative Word-classes are given the Analyser decides which is compatible with the Word-classes of the rest of the phrase. For example, if the word 'German' in the phrase 'German politics' has been assigned the alternative Word-classes 'N' and 'Adj', the Analyser eliminates the 'N' alternative because the grammar accepts only phrases of the form 'Adj+N' and not 'N+N'. 'Information retrieval' (if not treated as a compound word) will be analysed as 'Adj+N' because the word 'information' is assigned the Word-class 'Adj' in addition to the Word-class 'N'. If the alternatives are both (all) compatible the Analyser provides alternative phrase-structure analyses. Thus 'fighting ships' is analysed as both 'Adj+N' and 'Gerund+N'.

Example:

$\begin{Bmatrix} 01,310,\text{Adj} \\ 01,310,\text{N} \end{Bmatrix}$ + 02,162,Adj + 03,253,N → 01,310,Adj + 02,162,Adj + 03,253,N

i.e. the syntax permits 'Adj + N' but not 'N + N'.

(5) The Concept numbers (together with the running numbers) are transferred to a temporary store.

Example:

01,310,Adj + 02,162,Adj + 03,253,N →
01,310
02,162
03,253

(6) The Concept Store is compiled by extracting from the Dictionary all entries with the same Concept numbers as those now listed in the temporary store.

Example:

Concept Store

01,310					
02,162	→ America	310,N	economy	162,N	
03,253	American	{310,Adj / 310,N}	economize	162,V	
			...		
	America's	310,Ngen	historic	253,Adj	
	Americanize	310,V	historical	253,Adj	
	United States	{310,N / 310,Adj}	historically	253,Adv	
			history	253,N	
	...		...		
	economic	162,Adj			
	economical	162,Adj			
	economically	162,Adv			

(7) After eliminating the Concept numbers:

Example:

01,310,Adj + 02,162,Adj + 03,253,N → 01,Adj + 02,Adj + 03,N

the resulting 'bare' phrase-structure (or phrase-structures) is compared with the contents of the Syntax Pool. When one member of a set of phrase-structures is found to match, the whole set is extracted.

Example:

	01,Adj + 02,Adj + 03,N (the match)
01,Adj + 02,Adj + 03,N →	01,Ngen + 02,Adj + 03,N
	02,Adj + 03,N + *of* + 01,N
	03,N + *of* + 01,Adj + 02,N
	etc.

An alternative method for providing syntactic transformations by automatic derivation from the phrase-structure is described in section 4.

(8) Adding the Concept numbers from the temporary store (by matching the running numbers) to the set of phrase-structures provided by the Syntax Pool gives a list of phrases semantically equivalent to that analysed by the Syntactic Analyser (i.e. the search request). This list is the Transformation Store.

Example:

Syntax Pool	*Temporary store*	*Transformation Store*
01,Adj + 02,Adj + 03,N	01,310	01,310,Adj + 02,162,Adj + 03,253,N
01,Ngen + 02,Adj + 03,N	02,162 →	01,310,Ngen + 02,162,Adj + 03,253,N
02,Adj + 03,N + *of* + 01,N	03,253	02,162,Adj + 03,253,N + *of* + 01,310,N
03,N + *of* + 01,Adj + 02,N		03,253,N + *of* + 01,310,Adj + 02,162,N

2.4 STAGE II: SEARCH

(9) Each word of the Catalogue (or rather of those parts appropriate for 'subject' searches, i.e. titles, subtitles, abstracts, text abstracts, etc.) is compared with the contents of the Concept Store. If found it is assigned the Concept number and Word-class for that entry.

Example:

The words of the title of a Catalogue entry 'An introduction to the history of the American economy' are compared in turn with the Concept Store (see above); not until 'history' is reached is there a match:

An	introduction	to	the	history	...
ϕ	ϕ	ϕ	ϕ	253,N	

(10) Any word found in the Concept Store has its Concept number and Word-class compared with the contents of the Transformation Store. If there is a match with the *first* unit (Concept number/Word-class) of any of the phrase-structures, this phrase is extracted and placed in a temporary store. When other phrases have the same initial unit then the temporary store will contain more than one phrase.

Example:

253,N matches the first unit of the phrase-structure 03,253,N + *of* + 01,310,Adj + 02,162,N. This phrase is transferred to the temporary store.

(11) The next word in the Catalogue is looked up in the Concept Store and assigned its Concept number and Word-class.

(12) The Concept number and Word-class are now compared with the next unit of the phrase (or phrases) in the temporary store. If there is a match (i.e. if the prediction is fulfilled) the next word in the Catalogue is looked up and compared, and so on. By reiterating this process whole phrases in the Catalogue can be matched with the contents of the Concept Store and the Transformation Store.

Example:

Catalogue entry form	*Concept Store*	*Transformation Store (less running numbers)*
American economic history	$\left\{\begin{matrix}310,\text{Adj}\\310,\text{N}\end{matrix}\right\}$;162,Adj;253,N	310,Adj + 162,Adj + 253,N
history of American economy	253,N;*of*;$\left\{\begin{matrix}310,\text{Adj}\\310,\text{N}\end{matrix}\right\}$;162,N	253,N + *of* + 310,Adj + 162,N
economic history of United States	162,Adj;253,N;*of*;$\left\{\begin{matrix}310,\text{Adj}\\310,\text{N}\end{matrix}\right\}$	162,Adj + 253,N + *of* + 310,N

If at any point a word does not fulfil the prediction the process is recommenced at (9) with the word immediately succeeding. The temporary store will again be blank.

(13) When a whole phrase of the Catalogue has been matched, as described in step (12), the entry in which it occurs is read out for perusal by the inquirer as a potentially relevant document. Of course the whole entry need not be read out, only enough for the identification of the document.

2.5 An essential feature of the system is the Concept Store. Without it every word in the Catalogue (a large file) would have to be looked up in the Dictionary (another large file) and, since all would be found, they would then have to be compared with the Transformation Store. The inclusion of the Concept Store automatically reduces the look-up processes to the minimum required.

2.6 As texts in the Catalogue are in natural language, only 'meaningful' phrases will occur; thus no match will be made with any 'meaningless' phrases generated in the Transformation Store, e.g. the transformation of 'N+*of*+N' to 'Ngen+N' would produce 'butter's pound' from 'pound of butter'. Thus the contents of the Syntax Pool do not have to specify whether particular transformations are applicable to certain words or groups of words and not to others. This permits a considerable simplification. It is, however, stressed that none of the transformational rules given here or later as illustrations are considered foolproof; a far more complex and sophisticated system both semantically and syntactically is needed than the crude system outlined in this article.

2.7 Certain features of QUANTRAS are found in the SMART retrieval system:[11] both assign words to concept numbers, those of SMART being ranked in hierarchies, and both use a device for syntactic transformation. However, in SMART the whole document is analysed and provided with a set of identifiers (descriptors) and with a set of relationships; also statistical techniques as well as linguistic methods are used in analysis.

As in all language data processing systems the major problems of QUANTRAS are those of dictionary compilation and of syntactic analysis.

3 SYNTAX

Automatic syntactic analysis has been tackled for many years by workers in machine translation. Various systems have been evolved, all capable of analysing quite complex sentences. It has been shown (Gross[4]) that despite their superficial disparity all these systems are based ultimately on 'context-free grammars'. One major defect of all 'context-free grammars' is their inability to decide which of two or more alternative analyses of a given sentence is correct within its context. In machine translation a decision must be taken and so this defect must be surmounted or minimized. In the predictive syntactic analyser at Harvard,[6] for example, a choice between alternatives is based on their respective probabilities. Fortunately, however, QUANTRAS does not require decisions to be made, only that the alternative analyses be discovered. Thus any syntactic analysis program capable of

providing alternative phrase-structure analyses could be employed by QUANTRAS. If alternatives are given for a particular search request the system assumes that both (all) are intended (see step (4) in paragraph 2.3). An inquirer framing a request ambiguously cannot expect the computer to 'guess' his intention.

In QUANTRAS phrase-structures derivable by the Syntactic Analyser are grouped in semantically equivalent sets (the Syntax Pool), all members of a set being transformations of each other. However, the number of phrase-structures in natural language is large and the Syntax Pool is likely to be bulky. As an alternative method it is possible to envisage a system for deriving transformations automatically. This system would operate not upon the surface structures of phrases (i.e. those derived by the Syntactic Analyser) but upon their 'kernel structures'.* Whereas 'John loves Mary' and 'Mary is loved by John' have different surface structures, namely 'N_1+Pres+Vt+N_2' and 'N_2+Pres+be+en+Vt+by+N_1', they have the same kernel structure—and this fact explains their semantic equivalence. The differing surface structures are generated from the kernel structure by transformational rules (Chomsky,[2] p. 128–47; and many others)—in this case by an active-passive transformation. A phrase-structure analysis incorporating appropriate transformational rules would provide a kernel structure upon which other transformational rules could be applied for the derivation of semantically equivalent phrase-structures (which need not, for QUANTRAS, be 'meaningful'—see 2.6). Thus instead of specifying (as in the Syntax Pool) that phrase-structure A is semantically equivalent to phrase-structure B, we provide transformational rules to derive both from a kernel structure C.

The following tentative system, intended only for an information retrieval system such as QUANTRAS, analyses phrases in terms of a kernel phrase and applies to it various transformational rules. It thus replaces steps (4) and (7) of Stage I as described above.

4.1 The search-request phrase is analysed according to an immediate-constituent grammar. On the Word-class memberships (Adj, N, V, etc.) provided by the Dictionary the following rules are applied (each one is repeated until no longer applicable before proceeding to the next).

(1) Articles are eliminated.

(2) Adjectival groups are formed. (The subscripts refer to Concept numbers).

i $Adj_a + C + Abj_b \rightarrow Abj_{ab}$

ii $Adj_a \rightarrow AJ_a$

iii $AJ_x + AJ_y \rightarrow AJ_{xy}$ (where 'y' refers to two or more Concept numbers)

(A symbol such as Adj_{abc} means that any member of the Concept num-

* In the terminology of transformational grammar these are called more precisely 'deep structures' (Chomsky,[2] p. 17).

bers 'a', 'b', or 'c' which is also assigned the Word-class 'Adj' may occur).

(3) Similarly noun groups are formed.

i $N_a + C + N_b \rightarrow N_{ab}$

ii $N_a \rightarrow NP_a$

iii $NP_x + NP_y \rightarrow NP_{xy}$ (where 'y' refers to two or more Concept numbers). From the groupings so formed are now produced kernel phrase-structures.

(4) $AJ^1_x + NP^2_y \rightarrow NP\,(AJ^1_x + NP^2_y)$

(5) $NP(AJ^1_x + NP^2_y) \rightarrow NP^2_y + PR + NP^1_x$

Brackets are used to ensure that the correct order of 'NP's is made: thus 'Japanese economic history' is analysed as $AJ^1 + AJ^2 + NP^3$ and transformed from $NP(AJ^1 + NP(AJ^2 + NP^3))$ to $NP(AJ^2 + NP^3) + PR + NP^1$ and then to $NP^3 + PR + NP^2 + PR + NP^1$ [i.e. history of the economy of Japan] and *not* transformed as $AJ^1 + NP^3 + PR + NP^2$ and $NP^3 + PR + NP^1 + PR + NP^2$ [i.e. history of Japan of the economy (!)].

The kernel phrase-structure is now subjected to the transformational rules. These may be applied any number of times to any derived structure.

(6) $NP^1_x + PR + NP^2_y \rightarrow NP(AJ^2_y + NP^1_x)$

(7) $NP^1_x + PR + NP^2_y \rightarrow NP^2_y + Colon + NP^1_x$

Resolution rules are applied to every transformation of the kernel phrase. (They mirror the immediate-constituent and standardization rules (1)–(4)).

(8) $NP(AJ^1_x + NP^2_y) \rightarrow AJ^1_x + NP^2_y$

(9) $AJ_x \rightarrow Adj_x$ (where 'x' may represent any number of Concept numbers).

(10) $NP_x \rightarrow N_x$

4.2 As an illustration of the procedure we give the analysis, transformations, and resolutions of the phrase 'the presidential election in America'.

the presidential election in America	*(rules)*	
$Art + Adj_a + N_b + PR_f + N_c$		*Dictionary scan*
$AJ_a + NP_b + PR_f + NP_c$	(1), (2), (3)	
$NP(AJ_a + NP_b) + PR_f + NP_c$	(4)	
$NP_b + PR + NP_a + PR_f + NP_c$	(5)	*Kernel phrase*
$NP_b + PR + NP_a + PR_f + NP_c$		*Transformation I*
$N_b + PR + N_a + PR_f + N_c$	(10)	*Resolution of Tr. I*
election of (the) president $\left\{\begin{matrix}in\\of\end{matrix}\right\}$ *America*		
$NP(AJ_a + NP_b) + PR_f + NP_c$	(6)	*Transformation II*
$Adj_a + N_b + PR_f + N_c$	(8), (9), (10)	*Resolution of Tr. II 2*
$\left\{\begin{matrix}presidential\\president's\end{matrix}\right\}$ *election* $\left\{\begin{matrix}in\\of\end{matrix}\right\}$ *America*		
$NP(AJ_c + NP(AJ_a + NP_b))$	(6), (6)	*Transformation III*
$Adj_c + Adj_a + N_b$	(8), (9), (10)	*Resolution of Tr. III*

American / America's { *presidential / president's* } *election*

$NP_b + PR + NP(AJ_c + NP_a)$ (6) *Transformation IV*

$N_b + PR + Adj_c + N_a$ (8), (9), (10) *Resolution of Tr. IV*

election of { *(the) American / America's* } *president*

The application of rule (7) would produce phrases such as 'America: the election of the president' (on Tr. I), 'President of America: the election' (on Tr. I), 'America: the presidential election' (on Tr. II), and 'the American president: the election' (on Tr. IV).

An illustration, in brief, of rules (2) i, (2) iii, and (3) i:

the economic, social and political history of China and Japan

$Art + Adj_a + Adj_b + C + Adj_c + N_d + PR + N_e + C + N_f$

$AJ_a + AJ_{bc} + NP_d + PR + NP_{ef}$ (1), (2) i, (2)ii, (3)i, (3)ii

$AJ_{abc} + NP_d + PR + NP_{ef}$ (2)iii

$NP(AJ_{abc} + NP_d) + PR + NP_{ef}$ (4)

$NP_d + PR + NP_{abc} + PR + NP_{ef}$ (5) *Kernel phrase*

$NP(AJ_{ef} + NP(AJ_{abc} + NP_d))$ (6), (6) *Transformation III*

$Adj_{ef} + Adj_{abc} + N_d$ (8), (9), (10) *Resolution of Tr. III*

Chinese / Japanese { *economic / social / political* } *history*

It will have been noted that as members of the Word-class 'Adj' the Dictionary might supply 'genitive noun' forms (e.g. President's, America's). Also that articles are disregarded during the scan of the Catalogue.

Obviously the transformational rules outlined here are far too incomplete: more research is required on the transformations of noun phrases (on the lines of Lees' paper,[7]), and also on the kinds of noun phrases occurring in search requests and in the titles, subtitles, content lists, etc. of documents.

5 SEMANTICS

Since QUANTRAS does not analyse the whole texts of documents (unlike SMART[11] and all automatic indexing systems) the Dictionary will not need to contain the whole vocabulary of a language but only that part required for the analysis and matching of search requests. As these will be mainly formulated as noun phrases it might well include few verbal forms. (This in itself probably eliminates many problems encountered in machine translation, particularly the difficulties caused by homonyms and polysemes and by words with more than one possible grammatical role.) Inevitably the Dictionary will not on occasions contain words used in search requests; the system need not be blocked, however, if they are treated as if their form were invariant in all grammatical roles and as if they had no

synonyms. At a later stage (or earlier if the search proved unsatisfactory) new words would be incorporated with full details in the Dictionary.

5.1 The chief problem in the compilation of the Dictionary is the grouping of synonyms into 'Concept numbers'. K. Sparck Jones[12] has demonstrated the formation of 'rows' of synonyms by using dictionary definitions when they suggest alternative words or place the word in contexts (as in O.E.D.). When one word refers to different denotata it appears in different 'rows' (all the words in one 'row' refer to one denotatum), thus 'activity' appears in the following:

action	activity	briskness	liveliness	animation	
activity	animation				
activity	liveliness	animation			
activity	animation	movement			
activity	briskness	quickness	liveliness	speed	etc.

Such rows might well be formed into larger synonym groups and Sparck Jones proposed the statistical method developed in the theory of clumps. But while some groups so formed were felt to be 'correct', other groups split rows felt intuitively to belong together. This suggests either that greater sophistication is needed in the statistical technique or that linguistic intuition is not a valid basis for judgment.

Homonyms and polysemes will obviously occur in different rows and the denotata of the rows may be quite unrelated. If all the words in these rows are treated as synonymous, then the indiscriminate selection of all for a search request will potentially retrieve a large number of irrelevant documents. Although the number is less than might be supposed, since 'the probability is relatively low that the wrong meaning for one of the terms will appear in text with the correct accompanying terms' (Swanson,[14] p. 263), the risk must be minimized by some means of discriminating between the rows. As one method Sparck Jones[12] proposes that first the 'semantic distance' between two words in a sentence is measured by finding the smallest number of steps from one to the other (two rows having a common member being one step apart) and then the rows selected in the first and last steps are taken as the correct groups of synonyms for the two words in this context. An alternative method might be to use a classification of denotata (one denotatum being represented by a 'row' of synonyms) and to find the shortest route along the trees of the classification.

5.2 But there is a far more potent argument for the inclusion in the Dictionary of a classification scheme. It is well known that many index-users find a classification of descriptors or a system of cross-reference helpful in the clarification of their search request and for directing them to descriptors of possibly related material. For this purpose, if for no other, it appears to be desirable that some organization of descriptors (or 'Concept numbers') is incorporated in any retrieval system.

As no agreement on a universal classification is likely, the best solution seems to be that proposed for Syntol:[3] the provision of 'different conceptual organizations of partially identical indexing vocabularies in overlapping fields. . . Each one will correspond to a particular scale of knowledge, more refined in some, coarser in others.' Users of QUANTRAS could bring into operation different classifications of Concept numbers if and when desired.

5.3 However, one very common use of a classification in an index, namely for generic and specific searches, may not be necessary if the search is made on uncondensed texts rather than on restricted sets of descriptors (see paragraph 1.5). If large portions of document texts are searched then the assumption has been that somewhere, if the document is relevant, a phrase will be found to match the search request or its semantic equivalent. The document itself may treat the subject as part of a more general survey (in this case the phrase may occur in the contents list or the index) or treat only one particular aspect of it (here the introduction may place the content of the document in its broader field.) In these cases the scan will retrieve documents related generically and specifically to the request without using a classification scheme.

5.4 To convert complete document texts into machine-readable form for inclusion in a computer-based Catalogue would obviously be extremely costly unless the document were received already in a suitable form. (The development of automatic reading devices may solve this problem.) Some selection must be made—selection, it must be stressed, is not the same as condensation into an abstract or a list of descriptors—but nevertheless some loss is inevitable and some relevant documents may be missed. However, indexers know by experience that certain portions of a text are more informative than others, e.g. introductions, conclusions, abstracts, and that the author's own summaries in the form of title, subtitle, chapter headings, etc. can sometimes be quite helpful—but undoubtedly more research is needed on the information content of specific parts of documents.

Not every search would require the scanning of the whole Catalogue file: users might specify that only titles and subtitles be searched. A 'deeper' search, which could include abstracts, indexes, and a classification scheme (see paragraph 5.2), might retrieve a large number of unwanted document references in some cases.

The inclusion of texts in foreign languages in the Catalogue would require the translation of search requests. This could be achieved by expanding the Concept numbers to include foreign language equivalents and the Syntax Pool to include transformations for other languages. Alternatively, separate Dictionaries and Syntax Pools could be provided for each language (thus permitting users to specify the languages of texts to be scanned) and interlingual transformational rules, perhaps on the lines of those developed by Tosh[15] for German and English, would perform 'translations' of kernel phrase-structures.

6.1 Many difficulties must be overcome before a natural language search in document selection is at all feasible but it is arguable that the problems are no greater than those in automatic indexing processes. Automatic indexing involves the semantic and syntactic analysis of whole texts and their condensation as descriptors, and also, if it is a fully automatic system, the semantic and syntactic analysis of search requests. QUANTRAS involves similar linguistic analysis but only of search requests, which are then expanded into the variety of equivalents necessary for text scanning. Both condensation and expansion demand thorough knowledge of the semantics and syntax of natural language—conceivably a smaller segment for QUANTRAS than for automatic indexing—therefore it is worth considering whether the arguments against indexing *per se* do not justify the investigation of alternative methods of document selection.

REFERENCES

1. ANTAL, LASZLO. *Questions of meaning.* The Hague, Mouton, 1963. (Janua linguarum, ser. minor, 27.)
2. CHOMSKY, NOAM. *Aspects of the theory of syntax.* Cambridge, MIT, 1965.
3. GARDIN, J. C. *Syntol.* New Brunswick, Rutgers Graduate School of Library Service, 1965.
4. GROSS, MAURICE. On the equivalence of models of language used in the fields of mechanical translation and information retrieval. *Information Storage and Retrieval,* vol. 2, no. 1, 1964, p. 43–57.
5. JONKER, F. *Indexing theory, indexing methods, and search devices.* New York, Scarecrow Press, 1964.
6. KUNO, S. *and* OETTINGER, A. G. Multiple-path syntactic analyzer. *In: Information Processing: Proceedings of IFIP Congress 62.* Amsterdam, North Holland, 1963, p. 306–11.
7. LEES, R. B. The grammar of English nominalizations. *International Journal of American Linguistics,* vol. 26, no. 3, pt 2, 1960.
8. LUHN, H. P. The automatic derivation of information retrieval encodements from machine-readable texts. *Information Retrieval and Machine Translation: Int. Conf. for Standards on a Common Language for Machine Searching and Translation,* pt 2, 1959, p. 1021–8.
9. MOOERS, C. N. Summary of lectures no. 1 and no. 2. Presented at NATO Advanced Study Institute on Automatic Document Analysis, Venice, 1963. (Preprint.)
10. MORRIS, C. W. Foundations of the theory of signs. *International Encyclopedia of Unified Science,* vol. 1, no. 2, Chicago, 1947.
11. SALTON, GERARD. Automatic phrase matching. *Readings in automatic language processing,* ed. D. G. Hays. New York, American Elsevier, 1966, p. 169–88.
12. SPARCK JONES, K. Experiments in semantic classification. *Mechanical Translation,* vol. 8, nos 3–4, 1965, p. 97–112.
13. STEVENS, B. E. *Automatic indexing: a state-of-the-art report.* Washington, National Bureau of Standards, 1965. (Monograph 91.)
14. SWANSON, D. R. The formulation of the retrieval problem. *In: Natural language and the computer,* ed. Paul L. Garvin. New York, McGraw-Hill, 1963, p. 253–67.
15. TOSH, WAYNE. *Syntactic translation.* The Hague, Mouton, 1965. (Janua linguarum, ser. minor, 37.)
16. VICKERY, B. C. *On retrieval system theory.* 2nd edn. London, Butterworths, 1965.

17. YNGVE, VICTOR. In defense of English. *In: Information retrieval and machine translation.* Based on the Int. Conf. for Standards on a Common Language for Machine Searching and Translation, Cleveland, Sept. 6–12, 1959. New York, Interscience Publishers, 1961. pt 2, p. 935–40.