# on some compound semantic units and their mechanical recognition in english abstracts

The problem of avoiding duplication of research efforts by rapidly imparting scientific knowledge not taught at school from the authors who produce this knowledge to the scientists who use it has become so acute and has received so much attention after World War II that it has given rise to what can be claimed to be a new multidisciplinary science, which only partly overlaps with traditional library science : "scientific documentation", or "the information sciences" as it is sometimes called. One of the areas in this field is concerned with the use of electronic computers and other machines to speed up and systematize the exchange of scientific information. Within "automatic documentation" lies the subfield to which the present study belongs : automatic subject recognition in scientific texts or, more briefly, "automatic indexing". Studies in automatic indexing aim at mechanizing a task usually performed by professional indexers working in documentation centres. This task consists in specifying the subjects of scientific documents in such a way that, whatever the language of these documents and whatever the terminological and other differences within the same language, documents on the same subjects can be classified together. The customers of a centre can then retrieve the texts easily, provided they use the same subject specifications – terms of a classification, or "descriptors" – as those used by the indexer.

In this paper (1), I wish to submit a linguistic approach (2) to the problems of automatic indexing, and to apply this approach to the analysis of a few tests, which are listed in appendix 1.

The following questions can serve as an informal presentation of these problems. How could a computer classify scientific texts according to their meaning ? More precisely, how can a computer be "taught" how to "understand" a contemporary variety of written scientific English and how to recognize the subjects of abstracts – i.e. summaries of scientific documents – written in this kind of English, or list the passages it cannot "understand" and thus cooperate in a progressive enrichment of its linguistic "culture" ? In the solution adopted and discussed in this paper, two basic operations are required of the computer : dictionary-lookup and substitution (3).

Thus, *dictionary construction* and *the formulation of a system of substitution rules* are regarded as the two main objectives in the desing of an automatic indexing system (4). By way of illustration, let us suppose that a computer has to analyze part 2 of a given corpus, and to recognize a subject CAR-OWNER in sentences like *"This car belongs to my neighbour"* (5), or a related subject in *"My neighbour has bought this car last year"*. Automatic indexing will be possible if, after examining part 1 of the corpus, we have succeeded

(i) in giving the computer a dictionary including lists such as :
- A : this man, my neighbour, the lady next door, our friend, etc.
- B' : owns, is the owner of / the proprietor of, etc.
- B" : is owned by, belongs to, is (A)'s property, etc.
- B"' : has ought / purchased, etc.
- C : this car, a coupé, etc ...

(ii) in programming the machine to specify when, and *in how many stages* it is expected to replace the members of these lists by certain symbols – e.g. Cl B" A2 for *"this car belongs to my neighbour"* – and these symbols in turn by CAR-OWNER, or by a shorter symbol submitted to further substitutions if this subject is part of a more complex one;

(iii) and in having the computer signal a sentence such as *"this car belongs to a series produced by ..."* as unanalyzed if it happens that the machine does not have *"a series ..."* in its memory or if the occurrence of this segment immediately after *"belongs to"* has not been observed in, and/or predicted from part 1 of the corpus and is therefore ruled out. Thus, it can be said that "the output of a computer constitutes the first objective means of verification available to natural-language researchers" (P.L. Garvin, 1963b : 367).

## SOME PRESUPPOSITIONS

The texts analyzed in this paper appear in a collection of abstracts dealing with scientific documentation (P.C. Janaske, 1962). The choice of such a corpus and my intention to design an automatic indexing system for these texts are based on a number of presuppositions. I assume that automatic indexing is *feasible*, in spite of what has been argued by some prominent logicians (e.g. J. Bar-Hillel), and that, once the system has been set up, the computer memory will only need some minor additions that do not affect the design of the system. I also assume that abstracts are *simpler* to analyze than longer texts, or non-scientific texts, e.g. journalese or literary.

My third assumption is that *a corpus-based study* – i.e. a study going out from actual discourse – can contribute to the solution of some important linguistic problems, in particular the problem of where to draw the line between what is systematic and what is 'messy' in language. My fourth and main assumption is discussed in the following section. I believe that in the present state of linguistics – and in particular of lexical and semantic studies – my presuppositions cannot be proved right or wrong.

## TOWARDS A 'STRATIFIED' LINGUISTIC ANALYSIS.

In a famous review, Fred W. Householder Jr. (1952 : 261) has said : *"The rigor of linguistics is strictly fake rigor in many instances (...). We demonstrate that it is possible to determine the phonemes and morphemes of a language entirely without reference to meaning (...) but in practice we all use meaning (...)"*. If this is true, as I believe it is, meaning has to be given primacy in linguistic analysis.

This, in my opinion, has been done most clearly by Sydney Lamb (1964 a, b) in his "stratificational" view of language, which proposes to account for language in terms of a hierarchy of *autonomous strata* or levels *dominated* by the 'sememic stratum', i.e. the semantic level, to use more familiar words. The autonomy of the strata presupposes that the description at one level need not be a continuation of the description at the level immediately superior, and that types and rules of "realization" relate the descriptions at the various levels to each other. As G.N. Leech (1967) has recently pointed out, this conception is especially important in that it contradicts a working hypothesis which most linguists take for granted :

that the description of meaning is a continuation of grammatical description, or, in other words, that semantic statements are *necessarily* to be made about elements identified in grammar. In fact, as G.N. Leech stresses, the very opposite can be argued. To account grammatically for the two sentences *'This carpet belongs to Mrs. Jones'* and *'Mrs. Jones owns this carpet'*, which have the same semantic description (since they are equivalent in meaning), we need two grammatical descriptions, and this "diversification" (one of the types of "representation" or "realization" defined by S. Lamb, 1964 a : 66) can only be explained by a lexical fact, the alternative *'owns'-'belongs to'*. Hence G.N. Leech's important conclusion that, in such cases at least, *"lexical choice determines grammatical choice"*.

Much of the present paper represents an attempt to contribute to a stratified, or "stratificational", linguistic description. Unfortunately, it is difficult to be more definite about this matter at this stage and I can only refer to S. Lamb's and G.N. Leech's conceptions of the "stratificational model" because, as most grammarians are not interested in an explicit formulation of the semantic presuppositions of their investigation and most semanticists do not deal with grammar, this matter is not ripe for a better formulation than that of these two authors, and because, as far as I know, the only monolingual investigations that aim at a stratified description of language and are concerned with an extensive amount of data are the studies in automatic indexing which will be referred to below : those of the French C.N.R.S. team and, to some extent, those made under G. Salton (1966 : 169 - 188) to develop the SMART retrieval system.

In the next chapter I wish to show that indexing can serve as a method for the elicitation of semantic units and that the units thus discovered are

compound, i.e. consist of more than one elementary unit of grammar. The following chapters will be concerned with the problems of interpreting and describing the "compound semantic units" whose existence has been established by indexing. What I call "compound semantic units" is similar to J. Belin-Milleron's "liaisons" (J. Belin-Milleron, 1942; 1944; 1951), to G.N. Leech's "predications" (G.N. Leech, 1965; 1967), to P.L. Garvin's "predications" (P.L. Garvin, 1963a), to J.O'Connor's "connection forms" (John O'Connor, 1965), and especially to what in natural language corresponds to SYNTOL "syntagmata" (see Ch. III). Besides, my discussion of compound semantic units and of the dropping of some of their constituents may throw light on what Zellig Harris (1965a : 12) calls, after O. Jespersen, the "exocentric constituents" of "sentences" (e. g. the subject of the sentence *'Whether the experiments succeeded interested observers')* and on the semantic basis of his analysis : for Z. Harris, "exocentric" constituents are "word-sequences (phrases) such that we cannot replace them by any word of a characterizing category contained in them and obtain thereby another sentence of the language".

## II. INDEXING AS AN EMPIRICAL BASIS FOR THE DISCOVERY OF COMPOUND SEMANTIC UNITS

### INDEXING

Indexing consists in relating scientific texts to each other, irrespective of the language in which they are written, on the basis of shared meanings or subjects expressed in them. Indexers (or machines) can adopt one of the

following two procedures. The first method consists in extracting – i.e. quoting – "representative" or meaningful words and passages from the texts – as one does when underlining important passages while reading a document – and in relating the extracted words and passages to each other in lieu of the integral texts.

In the second method, indexing can be regarded as a kind of translation from the original texts into one or more terms – called "descriptors" – of a classification. "Descriptors" can be described as labels standing for a given subject, and a classification as a system that displays relations in one (or more) subject fields between the subjects symbolized by the descriptors. Here, I have adopted the second conception of indexing, and this study can be described provisionally as an attempt to explicit for a machine the operations which human indexers perform intuitively when they 'translate' scientific texts into descriptors of a classification. More generally, I shall go out from a *reader's* point of view; this implies that I shall not be concerned with what the writers of the texts 'thought' or 'intended'.

I say that "classifications display relations between subjects" because I do not wish to commit myself to a more definite discussion of classification systems. Like all deductive systems that claim to be empirically adequate, classifications are very difficult to describe. In spite of the great amount of work already devoted to the problem, it can be said that classification research is far from advanced and that the methods so far proposed to describe classifications are merely tentative (see J.C. Gardin, 1966; for a state-of-the-art discussion, see Baxendale, 1966); classifications are at best only working hypotheses about the meanings of texts in one or more scientific disciplines. The classification system used in the present study has been specially developped for the field of documentation dealt with in the texts I wish to analyze, and it is presented in the next section.

## A LIST OF DESCRIPTORS FOR THE LITERATURE OF DOCUMENTATION

The originality of the list of descriptors used here (F. Lévy) lies in the fact that, in spite of what its title may suggest, it is ***not*** a mere list of subject labels. The relations between the subjects for which the descriptors stand are defined in two ways :

(i) By the place occupied by each descriptor in the chapters, sections and subsections of the classification. The divisions into chapters and sections correspond to broad empirical distinctions that can be made in the field of documentation : for instance, the first chapter consists of descriptors about "scientific information" in general, exclusive of the practical (chapter II) and theoretical (chapter III) problems of automatic documentation. Special chapters also include descriptors for
- specifically linguistic studies that are only indirectly relevant to documentation;
- specifically mathematical and logical studies;
- studies in machine translation;
- and studies on the various types of machines used in documentation.

Further subdivisions – subsections and descriptors – correspond to the distinctions to be made between the general subjects just mentioned and more specific ones.

(ii) Each subdivision of the list of descriptors – including the descriptors themselves – is also defined by a *verbal* definition. The verbal definition used in this paper will be given below, in the next section.

The purpose of the list of descriptors is to make as many distinctions as are needed to obtain empirically adequate categorizations of the books and papers in the field of documentation. Consider, for instance, the word *"library"*. In everyday language, such a word may be perfectly unambiguous just as, say, "diabetes" would be for most non-medical speakers. In the field of documentation, however, the term is inadequate as a descriptor and at least three distinctions have to be made :

(i) If, for instance, the text to be indexed deals with the mechanization of some library tasks, it will have to be classified under a heading *mechanization*, and a descriptor of Ch. II or III will be used.

(ii) If a text or passage deals with, say, the production of indexes at the Library of Congress, this means that some kind of information processing is involved, and not merely the storage of documents. The list of descriptors therefore provides, next to the descriptor library, a descriptor *documentation centre* to express this distinction.

(iii) For a text on the organization of libraries in the United States, the descriptor will be *Networks of institutions dealing with scientific information*, or, more briefly, *scientific information networks.*

To explain the empirical bases of such distinctions, one could say that a text on *"U.S. libraries"* does not refer to the same 'things' as a text on *"the Library of Congress"*; to put it in a way that is more relevant to the purpose of this paper, the words and sentences will differ from one text to the other and the overlap – if any – will be small : e.g. a passage from the text on *"U.S. libraries"* that would be devoted to *"the Library of Congress"*.

## A SUBJECT, AND ITS EXPRESSION IN A FEW TEXTS

In this paper, I wish to analyze a few abstracts of my corpus that deal with the subject mentioned above : *"scientific information networks"*, henceforth SC. INF. NTWK. In the list of descriptors just presented, this descriptor appears in Ch. I (*"scientific information"*), section 2 (*"information needs and means"*), and it is defined as follows : "*General studies on regional, national, or international networks of scientific information institutions, their users, methods, and contribution to the information needs in one or more scientific disciplines*" (6).

After indexing the first 38 abstracts of my corpus, it appeared that 8 of them as a whole, and one of them partially (Abstract 38), could be assigned the descriptor SC. INF. NTWK.; the nine texts identified in this way are listed in appendix 1. Besides it appeared that, in these texts, a further distinction could be made between the segments that serve to *express* the subject and those that serve to *introduce* it.

For instance, in the following passage from Abstract 38, "The authors recount the origins and development of *documentation*, and its present organization *abroad and in Rumania*, especially *in science and technical subjects*", only the underlined words express the subject "SC. INF. NTWK"; similar passages in the other texts under investigation are also underlined in appendix 1.

The first point to be stressed here is that indexing is a purely *linguistic* operation, since it involves verbal definitions and texts : I have not *seen* the Rumanian documentation centres, but this does not prevent me from indexing the passage quoted above.

My second point is that indexing shows the existence of multi-word units – which I propose to call "compound semantic units". Their interpretation will be discussed in the following chapter.

In the interpretation I wish to propose, I shall not posit – as most semanticists do : (i) that there is a dictionary meaning, (ii) that the "effect" of discourse (U. Weinreich, 1963) is somehow to make us select part of this meaning, and (iii) that we perform such operations as "disambiguization" to select the right meaning. In the following chapter, I shall simply try to account for the empirical facts elicited by indexing as they are displayed in appendix 1.

## III. A BASIS FOR THE INTERPRETATION OF COMPOUND SEMANTIC UNITS

### PARADIGMATIC AND SYNTAGMATIC RELATIONS.

Meaning is primarily a psychological fact, i.e. a fact of individual consciousnesses; as such, however, it is neither observable, nor of course perceptible by a computer.

Meaning, then, can only be described in one of the two following ways. First, one can try to define the meaning of linguistic items in terms of the nonlinguistic facts – 'things', 'situations', etc. – they refer to. This approach, however, is impracticable here : unless one is satisfied with photography, there seems to exist no method or system other than language for

describing what in 'the world' corresponds to such expressions as *"documentation in Rumania"*. The other approach to semantic analysis, which will be adopted here, consists in defining the meaning of linguistic items in terms of 'relations' that they contract with other linguistic items; this, incidentally, is precisely what is required in the present study since I wish to account for the fact that linguistic items have been related to each other by indexing.

What this approach implies is best explained in the terms used by J. Lyons (1963 : 59) : "I consider that the theory of meaning will be more solidly based if the meaning of a given linguistic unit is defined to be the set of (paradigmatic) relations that the unit in question contracts with other units of the language (in the context or contexts in which it occurs), without any attempt being made to set up "contents" for these units"; for instance, "*a* is not synonymous with *b* because of its meaning, the fact of their synonymy is part of their meaning". In his study of Plato's vocabulary of knowledge, J. Lyons is not only concerned with well-known relations such as *synonymy* and *antonymy*; he also makes use of a *"consequence"* relation (e.g. between to know and to learn), and of a *"hyponymy"* relation. This is the relation that holds between a generic term (e.g. *flower*) and the corresponding specific terms (e.g. *tulip*) (7). Besides, J. Lyons's study carries a suggestion – which is confirmed by the present inquiry – that *hyponymy* may be the main "meaning relation", in the sense that it permits to account for more cases than the others.

J. Lyons's "meaning relations" will be used in the following section to analyze the passages that express the subject under investigation in this paper (SC. INF. NTWK.). On the other hand, J. Lyons's theoretical framework is not sufficient by itself to account for the "compound semantic

units" discovered by indexing so that it will have to be completed : J. Lyons's semantic investigation is only concerned with elementary units of vocabulary ("lexemes") and with "paradigmatic" relations between these units. In what follows, I wish to discuss a theory into which I propose to incorporate J. Lyons's "meaning relations" but which, I suggest, enables to account besides for "syntagmatic" relations, i.e. for the semantic relations between linguistic items *in* discourse. This theory is to be found in SYNTOL (Syntagmatic Organization Language : R.C. Cros *et al,* 1964), a model originally developed for the storage and retrieval of combinations of "descriptors" in computers. It should be noted that only those features of SYNTOL that are relevant to this discussion will be presented.

The basic framework offered by SYNTOL is rather simple, though more elaborate than that of comparable theories.

(i) SYNTOL proposes to classify lexical items under four categories within which relations such as J. Lyons's "meaning relations" can specify the paradigmatic organization of the vocabulary :
- *"Predicates"* (noted Q in the ensuing discussion) is the category of items that generally (e.g. *"especially"*, in my texts), or at least in a given subject field, have no autonomous meaning. The items that have autonomous meaning fall under one of the following categories.
- *"Entities"* (noted E) : e.g. *"child"*;
- *"States"* (noted S) for items expressing a process "that does not affect its object", or "the result of a process" (8) : e.g. *"idealization"* (of the mother by her child).
- *"Actions"*, noted A.

(ii) Besides, SYNTOL proposes to represent the subjects of scientific documents by pairs – called "syntagmata" – and triples of lexical items linked together by a *"relation"*. Only two of the SYNTOL "relations" are relevant to the present discussion :
- the *"associative relation"* (which I shall write *R 1*) which expresses 'qualification' or 'specification', in the most general sense of these words : e.g. an "entity" can be qualified by another, a "state" by an "action", etc (9). The orientation of this relation is fixed by convention : e.g. from a qualifying to a qualified "entity";
- the *"consecutive relation"* (which I shall write *R 2*) which can be roughly paraphrased by the English verb *"to affect"*. It corresponds to the relation E. Nida (1964 : **208)** has in mind when he describes the "N" of some "N of N" expressions as a "causative subject" : e.g. *"God of peace"* = *God who gives peace.*

The main implications of this system are best described by stressing a few original features of SYNTOL as compared to other schemes recently proposed by linguists such as E. Nida (1964), S. Lamb (1964 a, b), and P.L. Garvin (1963 a).

Whereas P.L. Garvin's (1963a) "predication types" are intended as *semantic* relations and are obtained by paraphrase (e.g. *a is followed by b* = predication type "succession"), the SYNTOL relations achieve a further reduction. It is only when this reduction leads to an ambiguous representation of the documents that some "*operators*" can be used : e.g. to indicate taht *b* is the "means" or "instrument" of an action *a*, rather than its "aim". In this paper, no "operators" were used because their absence did not seem to give rise to ambiguous descriptions.

The "State" category does not seem to have been posited by other authors than those of SYNTOL.

Besides, unlike Nida and Lamb, the authors of SYNTOL do not distinguish between the "agent" and the "goal" relations, except by conventionally orienting the *"agent"* to the *"action"* and the *"action"* to its *"goal"*. Both are represented as *R 1*, the associative relation, and the same relation is used when an *"Entity"* qualifies another (10).

The authors of SYNTOL stress the *axiomatic* character of their relations and categories. They do not claim, for instance, that they are based on linguistic considerations, e.g. "on the manner in which lexical items function in transformations" (E. Nida, 1964 : 62). They merely insist that the description is based on *pragmatic* considerations : the representation of a document is regarded as adequate when it enables to distinguish a document from the others *in the subject field considered.*

At last original feature I wish to stress in the attempt made in SYNTOL to formulate rules for *implicitation*, round the concept of "shunting" (11). These rules are concerned with the possibility of deriving "a *R 2c*" expressions from "a *R1bR1c*" expressions, *and* with the empirical adequacy of such rules : e.g. the derivation of the expression "vitamins *have an effect on* the skin" from "Vitamins *R 1* regenerate *R 1* the skin". The authors of SYNTOL do not claim universality for their "shunting" rules and present them as valid only for the medical field on which they have concentrated. For this reason, the use of "shunting" in this paper will be discussed after the analysis presented in the following section. All I wish to say here is that I shall use the "State" (S) category – which by definition is for items representing processes that do not affect their object – in cases when I wish to rule out "shunting".

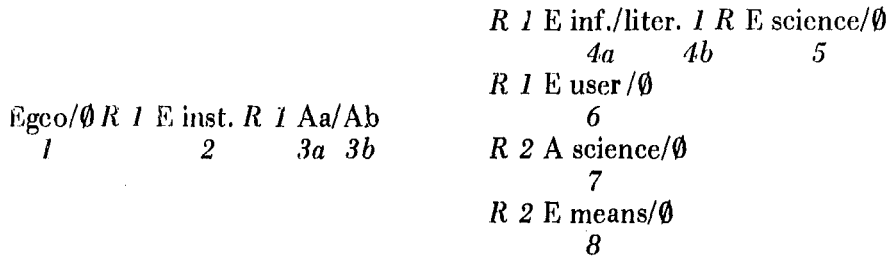
In the following section, I shall present an analysis based on both J. Lyons's "meaning relations" and on SYNTOL. After that, I shall discuss possible uses of the "shunting" operation in the present study, and of a partly related operation called "junction" ("raccordement" : A. Borillo *et al.*, 1966).

ANALYSIS OF SC. INF. NTWK.

In set-theory, a basic distinction is made between two complementary methods for describing "sets" of "ordered pairs" ("triples", ... "n-tuples") : the *tabulation method*, which consists in enumerating or listing the individual elements that "satisfy" an "ordering relation" – e.g. (2, 3), (4, 5), etc. for a relation ... *is less than* ... – and the *defining - property method,* which consists in stating one property (e.g. *X is less than Y*) common to all the elements belonging to the sed considered (S. Selby and L. Sweet, 1963 : 80). By analogy with this set - theoretical distinction, I propose to regard a description such as the one given below for the subject SC. INF. NTWK. in terms of generic words and expressions (J. Lyons's "hyponyms") classified according to SYNTOL categories and related to each other by SYNTOL relations as a *semantic property,* and the corresponding lists of individual items as the description of a *semantic set* by the "tabulation method".

In the statement of the "semantic property" of SC. INF. NTWK. given below, the segments labelled Q ("Predicates") have been left out to simplify the presentation. The SYNTOL relations are written *R 1 (R 2)* when oriented 'from left to right', and *1 R* in the opposite case. "Entities" and "Actions" are written respectively E and A. A bar (/) indicates an alternative, and /∅ means that the element at the left of the bar is droppable

without "shunting" (see next section). The generic terms or "hyponyms" are abbreviated ; e.g. geo = geographical.




When necessary, the items listed in what follows are given with the word or words that correspond to the SYNTOL relation relating the items considered to items of other lists : e.g. *"in"* that relates some items of the "geographical" list (e.g. *"Canada"*) to items of other lists. To simplify the notation I propose to use the underlined figures instead of the full notation : e.g. *1* instead of "E geo/Ø *R 1*". The heads of the lists and their synonyms are underlined by a continuous line (e.g. *institutions dealing with* for list *2*), and the heads of the sublists and their synonyms by a broken line, unless they coincide with the heads of the main lists. The sublists are noted *1'*, *1''*, *1'''*, etc., and this notation corresponds to what I have interpreted as increasing specificity. The only instance of antonymy is noted ≠ : abroad ≠ national. No instance of a "consequence" relation has been found. Some items are given merely because of graphic variations : e.g. United States, U.S.A., U.S.

*1'* international; in the NATO countries; for Africa, the Americas, Asia, Australasia, Europe, the Near East, abroad ≠

| | |
|---|---|
| *1"* | national, for each country, this country's; in Canada, Federal Republic of Germany, Federal, German, in the German Democratic Republic, France, French, Holland, Italian, Italy, U.S.A., in the United States, U.S., United States, in Rumania, Soviet, Swiss, of Switzerland; |
| *1"'* | geographical location, in the pertinent geographical region; province/state; city within province/state; the address ... for |
| *2'* | *institutions dealing with; scientific institutions responsible for; service bureaus/centers; centers; services;* government, Government; |
| *2"* | center, service; |
| *(3* | *the behavior of; 's role in; 's... activities; the ... activities of; efforts; systems of; system ... based on.* (12) |
| *3 a'* | *the dissemination of (... among); communications;* |
| *3 a"* | processing; |
| *4 a'* | *information; data* |
| *3 a' - 4 a'* | *documentation* |
| *3 b" - 4 b'* | photocopying; microfilm |
| *4 b'* | *holdings; collections; coverage provided over ...* |
| *4 b"* | literature |
| *4 b"'* | *1"* + periodical (13) |
| *(2 - 3 - 4* | *situation)* |
| *2' - 3b' - 4b'* | *libraries;* archives. |
| *5'* | *scientific; recorded; technical; in science and technical subjects; subjects;* |
| *5"* | subject of ... interest to; for each scientific or technical discipline; aeronautical |
| *6'* | *user.* |

| | |
|---|---|
| *6"* | *1"* + scientists and engineers; |
| *7'* | *achieve ... improvement in scientific productivity; recognized as ... part of research and development.* |
| *8'* | personnel; classification systems used; card-indexes. |

From an inspection of the texts, it appears that the following two *provisional rules* can be formulated for the expression of the subject SC. INF. NTWK. These rules represent combinations that are both 'minimum' and 'exocentric' in the sense that, if one of their constituents were dropped. the corresponding sequence of words would no longer express the subject SC. INF. NTWK; the examples quoted are from texts in which only these 'minimum' combinations occur :

*Rule 1 : 1 3 4*
*Example : "documentation in the (...) NATO countries"*
*3 - 4* *1*
(Abstract 3 1)

*Rule 2 : 2 3 4*
*Example : Scientific institutions ... responsible for the dissemination of ...*
*2* *3*
*information*
*4*
(Abstract 2)

It seems that there are two types of problems which the lists and rules given above leave unsolved.

(i) The *syntactic* behaviour of the items in the lists is not specified : this

would create serious difficulties if a computer had to look up a dictionary consisting of just these lists. In list *1*, for instance, we have an adverb (*abroad*), adjectives (*Swiss*), nouns (*of Switzerland*) and nominal groups (*in the German Democratic Republic*) introduced by prepositions. All these items behave differently when they qualify a noun like *documentation*. The SYNTOL relation is "realized" (S. Lamb) by the position of the item at the left of what it qualifies in the case of *Swiss*, by its position at the right in the case of *abroad*, by *in* or *of* and a lexical variation in the other two cases. The next chapter (Ch. IV) is devoted to such lexico-syntactic problems.

(ii) The unsolved *semantic* problems are both paradigmatic and syntagmatic. The paradigmatic problems concern the relations in the texts between the items of different sublists and inside the sublists. The syntagmatic problems mainly concern the possibility of dropping or implicitation of certain elements in the texts. The following section will propose some answers to these problems.

## "SHUNTING" and "JUNCTION"

As suggested above, SYNTOL proposes a "shunting" operation which permits to derive "a *R 2* c" expressions from "a *R1* b *R 1* c" expressions, and also from "a *R2* b *R2* c" expressions. In the latter case I propose to speak of "multiple shunting", In this section, I wish to consider successively the cases in which "shunting" is to be ruled out (i), and those in which "shunting" can be applied (ii).

(ii) "Shunting" will be ruled out when a *"b"* of the above formulas is classified as a "State" (noted S). This can be explained by presenting the semantic properties of the segments that serve to *introduce* the subject SC. INF. NTWK. The lists of items that correspond to these properties are not given for brevity's sake. To adopt the same king of conventions as in the preceding section, I shall write :

- 1 for "E author/paper/Ø *R 1*"
- 2a for "S discuss *R 1*"
- 2b for "S present *R 1*"
- 3 for "S reasoning *R 1*" (14)
- 4 for "E proper noun *R 1*"
- 5 for "A developement *R 1*"
- 6 for "A program *R 2*" (e.g. "program for improving")

All these categories except 2b are represented in the following simplified example from Abstract 2 :

This reviews the reasoning behind the desing by the National Science
1 2a 3 5 4
Foundation of a project to affect". The other possible combinations
6
are the following, with occasional dropping of 1 :

- 2 a 6 : "A proposed program to remedy ...
2a 6
ills"
- 1 2 a 3 5 : "The paper includes discussion
1 2a
of problems ... encountered in
3
the development of ..." (ab-
5
stract 29).

- 2 a 5 : "A proposal for the establishment
2 a 5
of" (Abstract 29)

- 2 a 2 a : "Report on a study of" (Abstract 2); "Summary of symposium on" (Abstract 31)

- 2 : "list of" (2 b : abstract 7); "discussion of" (2 a : abstract 6)

In the interpretation I submit, the ruling out of "shunting" when an "S" ("State") intervenes means that when an author "discusses" or "presents" something, he does not "affect" or "improve" the things discussed or presented; and that a discussion or presentation of "the reasoning behind" (or "problems encountered in") something does not "affect" this thing.

(ii) I now wish to suggest that "shunting" permits to account for some of the expressions of the subject SC. INF. NTWK. that occur in the texts under investigation (15), and thus enables to improve on the rules given for the expression of this subject. First of all, "shunting" describes the difference in meaning with which – as noted above – E. Nida is concerned in his analysis of "N of N" expressions : e.g. the indirect or "causative" relation between *"God"* and *"peace"* in *"God of peace"* (= God *who gives* peace) *versus* the 'direct' relation in *"Man of sin"* (= one who sins). In the present study, it permits to interpret expressions like *"a center for each scientific or technical discipline"* (Abstract 29), *"the subject of (...) interest ot the center"* (Abstract 7), *"institution, subjects"* (Abstract 25), as the result of a double "shunting", i.e. as a compression of "center that *handles information* on a subject", and *not* as parallel to, say, *'medical center'* in the

sense of *'hospital'*. I propose to write *R2* between brackets when "shunting" is multiple : e.g. (*R2*) when it is double, ((*R2*)), etc. Other examples of "shunting" are : the dropping of *2* in the shift from "... documentation *centers* of Switzerland" to "Swiss documentation" (Abstract 25), in "... documentation in the ... NATO countries" (Abstract 31), and in "documentation ... abroad ..." (Abstract 38); the dropping of *2-3a* in "this country's ... information" (Abstract 6), and in "international ... information" (Abstract 29).

*

* *

Three types of "*junction*" can be distinguished.

(i) The first type of "junction" is a complement of "shunting". It enables the machine to *explicit* "shunted" expressions every time the text to be indexed provides the elements needed. This can be illustrated by the following 'addition' applicable to the title of Abstract 25 quoted above :

| | | | | |
|---|---|---|---|---|
| | "of Switzerland" | *R1* "centers" | *R1* | "documentation" |
| + | "Swiss' | *R2* | | "documentation" |
| = | "Swiss" | *R1* "centers" | *R1* | "documentation" |

(ii) The second type of "junction" is similar to the first but it involves terms that are droppable *without* "shunting", i.e. the terms symbolized by ∅ in what precedes. It also involves elements which the authors of the texts leave implicit to avoid repetitions. The following

'addition' of the title to the only sentence of abstract 30 can serve as an illustration. The word order has been adapted for the present purpose.

| | | | |
|---|---|---|---|
| *Title* : Egeo = ∅ | *R1* "service centers" | *R1* "data processing" |
| + *Sentence* : "in Canada" etc. | *R1* "service bureaus" | *R1* ∅ |
| = "in Canada" etc. | *R1* "service centers/ bureaus" | *R1* "data processing" |

(iii) The third type is 'paradigmatic' in that it concerns the 'addition of specific terms to more generic terms, irrespective of their relative place in the texts.

The first case I wish to discuss is that of the items that have been bracketed in the lists presented above : (*3 : the behavior of,* etc.) and *(2-3-4 : situation).* In *"the behavior of scientific institution... responsible for the dissemination of recorded information"* (Abstract 2) for instance, I propose to regard *"the behavior of"* as an 'empty slot' so that, by 'adding up' this element (*3*) and *"the dissemination of" (3 a'),* we obtain the result *3 a'.*

Similarly, *"the ... situation in the United States"* would be deleted when 'added' to *"United States scientific and technical information services"* (Abstract 6), except for *3* (A *R1*) which would be preserved between *"services"* and *"information"*. If there is no element in the text to specify *3,* it can be translated by *R2*, the "consecutive" relation of SYNTOL. The 'adding up' of two or more synonyms must of course yield a fusion of them : e.g. *"service bureaus"* + *"service centers" (2'* : see Abstract 30).

In the other cases, however, it seems that two or more elements added to each other must be left distinct. The following examples from list *1* (geo) may serve to illustrate this point.

- "(documentation centers) *in the German Democratic Republic*" + "*The address ... for* (82 documentation centers)" (Abstract 7).
- "(documentation centers) *of Switzerland*" + "(documentation centers) *... geographical location*" (Abstract 25)
- "(documentation) *in the ... NATO countries*" + "(efforts) *in Canada, France, Holland, Federal Republic of Germany, U.S.A., and Italy*" (from which a NATO country such as Belgium is left out) (Abstract 31).

## IV. A BASIS FOR SYNTACTIC DESCRIPTION.

### ADAPTING "STRING ANALYSIS" TO SEMANTIC INTERPRETATION AND TO LEXICAL FACTS.

As suggested at the beginning of this paper, the two basic operations which, in my opinion, are required of a computer in automatic indexing are *table-lookup* and *substitution.* The semantic interpretation submitted in the preceding chapter and the corresponding notations which were proposed represent a first step towards the preparation of these two operations, but this is not enough for automatic processing because the syntactic behaviour of the constituents of compound semantic units is not specified, and because, unless programmed to do otherwise, a computer can only handle words or other symbols separated by spacing.

As will be argued in this chapter, Z. Harris's "String Analysis of Sentence Structure" (1965a) can be adapted and extended to specify the syntactic behaviour of the constituents of compound semantic unitc. "String analysis" is basically an attempt to write an 'algebraic' grammar of English. The possibilities which ideally the system offers are described by Z. Harris (1965a : 28) as follows : "When each segment of a formula *F* is replaced by a word which is a member of the category occupying that segment, the result is a word-sequence which occurs in sentences of the language precisely af *F* occurs in the string-formulas of the grammar". The basic features of Harris's analysis can be summed up as follows :

(i) Harris posits a few "sentence types" including a main "center string" *Subject Verb Object* in which *S V O* can have various "values".
(ii) The values of S and O are stated by means of symbols such as N (noun), A (adjective), P (preposition), D (adverb), or of individual words (e.g. "whether").
(iii) The place of the words corresponding to the symbols N, etc., is specified by the order of these symbols in the formulas, or by special notations.
(iv) S and O can themselves be a "center string", or an "exocentric" string (see Ch. I).
(v) O can have the value zero.
(vi) What is not "center string" is categorized as "adjunct", and is also defined by symbols such as N, A, P, etc., and strings of such symbols. The main categories of "adjuncts" are : adjuncts of P, of D, of A, of N, of V, of "center strings", and adjuncts introduced by conjunctions (K).
(vii) *The members of the general syntactic categories represented by the symbols listed above are not specified (i.e. listed). They are merely*

*represented by 'dummy' subscripts and notations,* and suggested by examples. For instance, for a subcategory of "left adjuncts of nouns" which is written "N of", Harris's notation (1965 : 36) simply means "left adjuncts of nouns" of a distinguishable type, and only the following comment is added : "a particular subcategory of *N* including *kind, type, sort : He is a sort of investigator*".

(viii) "String analysis" can be extended into a system for stating *semantic* relations such as equivalence, as Z. Harris himself seems to suggest (Z. Harris, 1957 : 297), *provided the members of the categories are listed* : for instance,
N1 of N2 = N2 of N1, e.g.
"This type of bacteria" = "Bacteria of this type",

(ix) Similarly, it can be adapted to serve as a system for *semantically-based substitutions, provided LEXICAL facts and categories are taken into account.* For instance, Z. Harris (1957 : 286) writes *"NPN = N"*, and I would suggest to write *"NPN = (N)"* to stress the fact that "=" means "can be replaced by". Thus *"This type of bacteria"* = N1 of N2 = (N1-2) = *"Bacteria of this type"* = N2 of N1. Similarly AN = (N), for instance. But, as noted above, *lexis* decides whether such substitutions are possible or not. Apart from the 'mixed' case illustrated above, Z. Harris (1957 : 297) also discusses a "PN = A" structure (e.g. lists of names are appended) and a "NP = A" structure (e.g. A number of boys were arguing) but, as Harris himself suggests, each structure has to be defined by a specific list of lexical items : *"Boys of a number"* is dubious, and if it is acceptable, it is not comparable to *"A number of boys"*.

In the following section I wish to present a few analyses based on the kind of adaptations of "string analysis" I have just proposed, *and* on the results of the semantic interpretation submitted in the preceding chapter (Ch III).

The results of semantic interpretation will yield both *semantic* (1, *1*, etc.) and *lexical* subcategories (e.g. "Swiss -" *versus* " - of Switzerland" = A*1*" - *versus* - PN*1*"), i.e. the *membership* of the syntactic categories symbolized in the formulas of "string analysis". Besides, semantic interpretation will permit to outline a variety of "string analysis" consisting of formulas which, for the purpose of this study at least, seem to account for a more satisfactory *segmentation* than some of the formulas proposed by Z. Harris.

## OUTLINE OF A MECHANIZED INDEXING SYSTEM,

The system outlined in this section consists of three types of rules corresponding to three operations : *simple substitutions, reductions,* and *normalizations* (16).

The *simple substitutions* consist of replacing each word by a syntactic and, if necessary, a semantic notation : e.g. N *234* for "libraries". Occasionally, more than one word may be replaced by a single notation : e.g. N*1* for "United States" (17).

At the basis of the *reduction rules* lie the following assumptions.

(i) The first one can be regarded as the main axion of string theory and of its computational extensions : discourse can be described in terms of "regularities of class combination" (Z. Harris), i.e. as repetitive cooccurrences of a limited number of syntactic structures consisting of a limited number of parts-of-speech categories. Consequently, long and complex strings can be broken down into sequences of

shorter strings to be processed first and to be considered as single elements in later scannings. instead of applying rules set up for a large number of elements in a single scan, we can have more than one pass and apply simpler rules involving a limited number of elements and better adapted to computation. The relative position of the items in discourse is specified by the order of the category-symbols in the string formulas.

(ii) The second assumption, which also forms an essential part of string theory, as I interpret it here, concerns the existence of three basic types of strings, and the possibility of reducing them to a single element. It can be expressed by the following axiomatic formulas, in which X stands for an element or string belonging to any kind of syntactic category and the equal sign for "can be replaced by or reduced to" :

- *X K X* = *X*, i.e. coordinated elements can be reduced to a single element;
- *X left X* = *X*, i.e. if X left is a left adjunct of X, the string *X left X* can be reduced to a single element;
- *X X right* = *X*, i.e. if X right is a right adjunct of X the string *X X right* can be reduced to a single element.

In other words, the three patterns considered – coordination, left and right adjunction – yield the same reduction. This implies that the reductions I wish to propose are independent of syntactic structure, i.e. that the reduced strings are assigned no syntactic description, in the form of bracketings or in the form of trees. Thus, the strings "*Swiss* (libraries, archieves)" and "(*historical* development), holdings" (Abstract 25) are both reduced to N, i.e. nominal group, irrespective of the different syntactic functions of "Swiss" and "historical" (18).

(iii) The third assumption can be put as follows. As a preliminary step towards mechanical recognition of meaning, the step-by-step reductions of strings of syntactic notations to a single notation can be performed *regardless of the semantic notations* attached to the syntactic notations.

The semantic notations just need to be preserved for independent treatment, such as rearrangement ("documentation centers" = N *34* N *2* = N *342* = N *234*) or deletion of invalid sequences ("nervous centers" = A *n'* N *n/2* = N *n'n/2* = N *n'n*). In this sense the reductions I wish to propose can be regarded as independent of meaning as they are of syntax. If my assumptions are correct, the proposed reductions can avoid the main difficulty of mechanized indexing :

- some strings have the same or a similar string formula but a different meaning, e.g. TNPN Ved/A P(N) for "*the behaviour of ... institutions ... responsible for (the dissemination of recorded information)*" and "*the construction of ... project supported by (the National Science Foundation)*";
  and some strings are syntactically different but semantically related : e.g. "*libraries*", "*documentation centers*", "*scientific and technical information services*" and "*scientific institutions that are responsible for the dissemination of recorded information*"; or "*historical development*" and "*origins and development*" (on this, see G. Salton, 1966). What I propose is to reduce all these strings to N, i.e. nominal group, accompanied by the appropriate semantic notations.

(iv) My last point is that *lexical* facts – facts which belong to the history of a particular language and society – determine which of the three

patterns mentioned above (coordination, left and right adjunction) actually occur and how they are related to each other. Consequently, *ad hoc lexical lists* are regarded as essential to ensure the delicacy of the rules. The rules consists of a formula *and* a lexical domain. If, for instance, we wished to replace the right adjuncts of "documentation" in documentation ... *abroad and in Rumania* (I) ... *in science and technical subjects* (5)" (Abstract 38) by equivalent left adjuncts such as those of "information" in "*United States* (1) *scientific and technical* (5) information" (Abstract 6) we would probably have to give up the left adjunction pattern contained in the right adjunct "in science and *technical* subjects", *and* use a partly different vocabulary : e.g. "The authors discuss Rumanian and non-Rumanian / foreign scientific and technical documentation".

The purpose of the *normalizations* I wish to propose is to simplify the reduction rules in two ways. First, I have tried to limit the number of different syntactic categories involved in them : "This reviews" is rewritten as N2 P, which is the notation for "review of", "discussion of", etc; "developed by" as P5. Besides, I have tried to limit the length of the string formulas to be reduced, for instance by deleting items such as *"which has been"* labelled Q (predicate) by the semantic interpretation. Thus "N *which has been developed by* N" is rewritten N P5 N. The intention is to arrive at a final representation of the abstracts under investigation in the form of a single *N-Symbol followed by a string of semantic notations.* The word "normalization" is used rather than the word "transformation" to stress the arbitrary character of the proposed rules, which – as suggested in Ch. I – do not give primacy to strictly syntactic facts as transformational theories do, like most other current theories of grammar : for instance, there is no reason for considering the nominal form as primary

apart from the fact that the nominal group is by far the best represented pattern in the abstracts considered.

The following example may serve to show how the normalization and reduction rules work :

"An annotated index to 459 of the most important Swiss libraries, archives, documentation centers, etc., arranged by geographical location." (Abstract 25).

Simple substitutions
T VedQ N2 Pto AQ Pof T D Q AQ A*1* N*234*

K N *234* K N *34* N *2* K etc. K Ved P by A*1* N*1*

Normalizations

| | | |
|---|---|---|
| An annotated index to | : T Ved Q + N2P | = Ø N2P |
| 459 of | : AQ Pof | = Ø |
| the most important | : TDQAQ | = Ø |
| , etc., | : K etc. K | = Ø |
| arranged by | : Ved P | = P |

Reductions

| | | |
|---|---|---|
| Swiss libraries | : A*1* N *234* | = N *1234* |
| documentations centers | : N *34* N *2* | = N *234* |
| geographical location | : A*1* N*1* | = N *1* |

The sentence can now be rewritten as follows :
N2PN*1234* K N *234* K N *234* P N*1*

By means of the rules NPN = N, and
NKN = N,
and by a few simple operations on the semantic notations we obtain the final representation of the sentence : N2 *1234*. From now on, as in the above examples, I have further simplified the simplified semantic notations proposed in Ch. III, to make the rules presented below more readable.

*

* *

The order in which the rules are given is intended to reflect a possible order of application of these rules. The rules cover practically all the data analyzed in this paper.

1. Simple substitutions

1.1. the/a/an/this/such a = T
1.2. and/or/:/;/,/, and = K
1.3. carried out = Ved 5
1.4. T + : National Science Foundation / Operations Research Group / Case Institute of Technology = N4
1.5. origins K development / historical development = N5
1.6. T/Ø + : U.S.A./U.S. United States/German Democratic Republic/ Americas/Near East/Federal Republic of Germany = N*1*
1.7. research K development/scientific productivity = N7
1.8. in addition to/in order to = P

The other substitutions need not be mentioned here as they concern single

word items, most of which are listed below under the relevant rules or have been listed in Ch. III.

2. Normalizations

2.1. Deletions of Q. They concern items labelled "predicates" in the semantic analysis. I assume provisionally that – unless otherwise specified – the deletions can be made without referring to the context.

2.2.1. Some of / each of / 459 (or another number) of (AP of); the reasons behind / the scope of / the character of / availability of (T/Ø NP); the most important (TDA); main / principal / initial / present / current / integral / primary / net / each / several / six-point / one / three / 82 (or another number) / pertinent / following (T/Ø + A / Ving); annotations / ills (T/Ø N); long range (AN); also / only (D); ,particularly/ ,especially (KD); in addition, (PNK); ,etc., (K etc. K); ) (end of parenthesis).

2.1.2 that / which / Ø + is / are / has / have + been / would / must + be + A / Ved / Ving = A / Ved / Ving.

2.1.3. T / Ø + alphabetical / annotated + N2P = N2P

2.2. Other deletions.

2.2.1. N6 + information = N6 (user information)

2.2.2. Here it seems necessary to introduce the negative category symbol

P, standing for non-preposition :

Ved P = Ø; e.g. "classification systems *used,*" and "Alphabetical indexes ... (are) *included.*"

2.3. *Phrases rewritten as N2 P.*

2.3.1. T/Ø V2 P : e.g. Lists, This reviews.

2.3.2. TN1 V2 : e.g. the author / paper / article / authors + discuss(es)/ recount(s).

2.3.3. TNIVT/Ø N2 P : e.g. the paper includes / presents a review of, the author presents discussion of.

2.3.4. K / P Ving T / ØN2 P : in addition to presenting a review of /, including description of

2.3.5. TN1 V2 that : for V2 = concludes

2.3.6. there would be

2.3.7. additional information concerns / descriptive data include.

2.4. *Ving 3 = N3* : for Ving 3 = planning / reasoning / thinking.

2.5. abroad = PN1

2.6. *Strings rewritten as P.*

2.6.1. PVNP : to determine ways to / to achieve improvement in

2.6.2. PV : for V = affect(ing) / improve / improving / remedy(ing),

2.6.3. PNP : of interest to / as part of. The difference with 1.8 (e.g. in order to) is that here insertions are possible : e.g. of *primary* interest to (see 2.1.1.).

2.6.4. A / Ved / Ving P : one for; written in / based on / arranged by / responsible for / dealing with / used in / encountered in / recognized as; developed by / carried out by / supported by (=P5); provided over; given for. In some cases, particularly the last two, the simplification attempted here may not hold as the prepositions "over" and "for" are in fact related to a noun preceding Ved.

2.6.5. ( , i.e. beginning of parenthesis.

2.6.6. 's = * P. The special notation is introduced to signal cases interpreted as left adjunction (see 3.2.4.).

2.6.7. K its, as in "documentation and its ... organization". A special notation could also be used here to express the fact that we have to do with a 'mirror image' of "N of N" : e.g. organization of documentation.

3. *Reductions*

3.1. *Coordinated pairs*

3.1.1. A5K A5 = A5 : scientific and / or technical.

3.1.2. N/Ving K N/Ving = N
N3 K N3 : reasoning and thinking
N5 K N5 : design and construction
N1 K N1 : province or state
N34 K Ving 34 : microfilm and photocopying
N6 K N6 : scientists and engineers.

The relative position of N or A may be reversible here, contrary to the general notational convention of string analysis.

3.2. *Left adjunction*

3.2.1. PP = P : for instance (for, i.e. parenthesis + for.

3.2.2. TN = N

3.2.3. T / Ø + A / Ved / N + N/Ving = N
Several passes with the same rule have to be envisaged here.

| | | |
|---|---|---|
| Pass 1 : | N1 N1 = N1 : | NATO countries |
| | A1 N1 = N1 : | geographical region / location |
| | N2 N2 = N2 : | service + bureaus / centers / institutions |
| | A5 N2 = N2 : | scientific / technical / aeronautical + services / bureaus / centers / institutions, or N2 obtained as above, |
| | N1/A1 N2 = N12 : | for N2 as above and for A1 = Federal / German / French / Italian / Soviet / Swiss / national / international; for N1 = U.S./U.S.A./United States. |

| | | |
|---|---|---|
| | A1 N2 = N12 : | Federal Government |
| | A1 / N1 N(3)4 = N 1(3)4 : | for A1/N1 in the two rules given above, and for N34 = documentation or N4 = information, e.g. Swiss documentation |
| | A1/N1 N234 = N1234 : | N 234 = libraries / archives : e.g. Swiss libraries |
| | N34 N2 = N234 : | e.g. documentation centers, (microfilm and photocopying) services. |
| | N5 N(3)4 = N(3)45 : | e.g. aeronautical documentation, scientific information. |
| | A5 N4 = N45 : | for N4 = communications, and for A5 = scientific / technical (?) |
| | Ved 5 N4 = N45 : | for Ved5 = recorded |
| | N4 Ving3 = N34 : | data processing |
| | A5 N5 = N5 | for N5 = subject(s) discipline(s); for instance, scientific discipline. |
| | N5 N5 = N5 : | periodical literature |
| | A1/N1 N6 = N6 : | e.g. U.S. (scientists and engineers) |
| Pass 2 : | A1 N5 = N5 : | e.g. Soviet (periodical literature). |
| | N45 N3 = N345 : | for N3 = behavior / activities / efforts / system(s) For instance, scientific information system. |
| | N45 N2 = N245 : | e.g. scientific and technical information services |

Pass 3 : A1/N1 + N345/N245 = N1(2/3)45
For instance, international technical information system

3.2.3. NPN = N. This seems to be restricted to the following cases (cp. however 2.1.1. : the scope of, etc.) :

- (Federal Government)'s N345/N245 = N12(3)45 : e.g. Federal Government's scientific and technical information activities
- (Federal Government)'s N3 for N3 = role (in ...)
- N1*P N(2/3)45 for N1 = country / Switzerland / etc. except U.S. etc. For instance : this country's scientific and technical information
- N1 P N234 for N1 = address / perhaps (geographical location) etc., i.e. specific N1 (N1'"). Example : the address (is given for) documentation centers.

3.3. *Coordination and right adjunction*

After the above reductions, the abstracts under investigation would only consist of coordinations and right adjunctions to be reduced by applying the following rules *recursively :*

3.3.1. N K N = N. The relative position of the N's is not specified by the formula since the terms seem to be reversible. It also appears to be useless to give this rule a lexical domain, for any series of N's, it seems, can form an enumeration : e.g. "institutions, subjects, and personnel" (Abstract 25), "Canada, France, Holland, Federal Republic of Germany, U.S.A., and Italy" (Abstract 31).

3.3.2. PN + K + PN = PN
Examples :
.....
P5 (= supported by) (N.S.F.) KP5 (= carried out by) (Abstract 2); by (province or state) K by city within (province or state) (Abstract 30). The qualifications of 3.3.1. also seem to apply here.

3.3.3. NPN = N. For this rule, I also assume provisionally that the relative position of the N's need not be stated explicitly.

Example : Abstract 2
.....
(Review) of (reasoning and thinking) behind (design and construction) of (project) PN (see 3.3.2.) at (Case Institute of Technology) P (= to determine ways to affect : see 2.6.1. and 2.6.2.) (behavior) of (scientific institutions) P (= that are responsible for) (dissemination) of (recorded information) P (= in order to achieve a net improvement in) (scientific productivity).

*

* *

Programming and computation would no doubt point to a number of weakness in the system outlined above and in the theoretical assumptions on which it rests. Before I can report on computational tests of the system, all I wish to do, by way of conclusion, is to suggest briefly how the delicacy and recognition power of the proposed rules could be increased

by enlarging their *lexical domain.* The extensions could be made on the basis of the meaning relations discussed in Ch. III and of the distinctions made by string analysis, particularly that between left and right adjunct lists. This can be illustrated by examples taken from two texts which have not been analyzed in this paper : abstracts 42 and 47 quoted in appendix 2. They are the first two abstracts coming after abstract 38 (appendix 1) in the corpus that deal with the subject SC. INF. NTWK analyzed in this paper. On the basis of the *hyponymy* relation (the relation between a generic item and the related specific items), and by analogy with compound semantic units analyzed in Ch. III a competent speaker of English could probably predict expressions like the following : "*medical* information / documentation" (Abstract 42) after "*aeronautical* documentation (Abstract 31); "medical information *in the Soviet Union*" (Abstract 42) and "documentation *in West Germany*" (Abstract 47) after "documentation centers *in the German Democratic Republic*" (Abstract 7); "*organizations engaged in medical documentation*" (Abstract 42) after "*institutions dealing with* documentation" (Abstract 38); or even types of adjuncts not discussed in this paper such as "documentary activity", ... *from the* professional and organizational *point of view*" and "... *at* national *level*" (Abstract 47). It may be noted in this connection that some predictable and perfectly valid expressions of SC. INF. NTWK. would have a very low probability of occurrence, because the reality they would apply to is improbable : e.g. "systems of oceanographic documentation centers of Lichtenstein". Extensions based on the consequence relation (and probably on other meaning relations) are also possible : "*better* dissemination of information" (Abstract 42), for instance, could be predicted from "*for improving* the dissemination of scientific information" (Abstract 5) – in which "for improving" has intentionally been reduced to P, just as "*better*" would be deleted as a Q for simplicity's sake. Such examples go to show

that, as it stands, the system outlined in this note could serve as a basis for mechanized subject recognition in texts not previously analyzed by an indexer.

J. NOËL
Research Student
F.N.R.S. and Canada Council,
Universities of Liège and
Victoria, B.C.

BIBLIOGRAPHY.

Phyllis BAXENDALE (1966) : *Content Analysis, Specification and Control,* in Carlos A. CUADRA, ed., *Annual Review of Information Science and Technology,* Vol. 1, New York, 1966, pp. 71-106.

J. BELIN - MILLERON (1942, 1944) : *La réforme de la connaissance,* Vol. I - III, Tours and Paris, 1942 and 1944.

J. GELIN - MILLERON (1951) : *De l'étude concrète des concepts politiques et moraux,* in *Revue de l'Institut de Sociologie,* 1951, no. 1, pp. 23-33.

A. BORILLO, N. SIOT-DECAUVILLE, J. VIRBEL (1966) : *Etudes sur l'indexation automatique, Phase de traduction syntaxique,* Rapport semestriel no. 2, December 1966, 46 p. (An internal report of the Section d'Automatique Documentaire, Marseilles).

M. COYAUD (1964) : Le SYNTOL (*Syntagmatic Organization Language*), *Etude d'un système général de documentation automatique,* Vol. IV *Analyse automatique,* Report EUR 423, f, Brussels, 1964, 79 p.

R.C. CROS, J.C. GARDIN, F. LEVY (1964) : *L'automatisation des recherches documentaires : un modèle général, "le SYNTOL",* Paris, 1964, 260 p. See an English version of this book by J.C. GARDIN (1965).

EUR (ATOM) 409 f. (1964) : *Etudes de documentation automatique, Analyse et comparaison des langages documentaires,* Vol. I, Brussels, 1964, 106 p.

J. C. GARDIN (1965) : *SYNTOL*, In Susan ARTANDI, ed., Rutgers Series on Systems for the Intellectual Organization of Information, Vol. II, New Brunswick, N.J., 1965, 106 p.

J. C. GARDIN (1966) : *Eléments d'un modèle pour la description des lexiques documentaires,* in *Bulletin des Bibliothèques de France,* Vol. II, no. 5, 1966, pp. 171-182.

Paul L. GARVIN (1963a) : *Research in Semantic Structure in Inductive Methods in Semantic Analysis,* Technical Note no. 6, Canoga Park, California, 15 January 1963, 22 p.

Paul L. GARVIN (1963b) : *Natural Language and the Computer,* New York, 1963, 398 p.

Zellig S. HARRIS (1957) : *Co-occurrence and Transformation in Linguistic Structure,* in *Language,* Vol. 33, no. 2, 1957, pp. 283-340.

Zellig S. HARRIS (1965a) : *String Analysis of Sentence Structure,* Third printing, The Hague, 1965, 70 p.

Zellig S. HARRIS (1965b) : *Transformational Theory,* in *Language,* Vol. 41, no. 3, 1965, p. 389.

R. HOFFMANN (1966) : *Programmes d'analyse en vue de l'indexation automatique de textes documentaires,* Memoir presented at the University of Liège, 1966.

Fred W. HOUSEHOLDER, Jr. (1952); Review of Zellig HARRIS, *Methods in Structural Linguistics,* in *International Journal of American Linguistics,* Vol. 18, no. 4, 1952, pp. 260-268.

Paul C. JANASKE, editor (1962) : *Information Handling and Science Information, A Selected Bibliography 1957-1961,* Washington, D.C., 1962.

Sydney M. LAMB (1964a) : *The Sememic Approach to Structural Semantics,* in A. KIMBALL ROMNEY and Roy GOODWIN D'ANDRADE, eds., *Transcultural Studies in Cognition, American Anthropologist,* 66, no. 3, part 2, 1964, pp. 57-78.

Sydney M. LAMB (1964b) : *On Alternation, Transformation, Realization, and Stratification,* in C.I.G.M. STUART, ed., *Monograph Series on Languages and Linguistics,* no. 17, 1964, pp. 105-122.

G. N. LEECH (1965) : *"Scale and Category" Semantics,* University College, London, 1965, 38 p. (Paper read at the Philological Society on 12th November, 1965).

G.N. LEECH (1967) : *Systems and Structures of Meaning – An Outline of a Semantic Theory,* pre-publication version of a book, kindly communicated by Professor Leech.

R. B. LEES (1960) : *The Grammar of English Nominalizations,* in *International Journal of American Linguistics,* Part II Vol. 26, no. 3, July 1960, p. 1-205; quoted in B. GRUNIG *Les théories transformationnelles,* Part II, in *La Linguistique,* Vol. I, no. 1, 1966, p. 93.

F. LEVY : *Information Processing of Non Numerical Data, Provisional List of Descriptors,* Groupe d'Etudes sur l'Information scientifique, S.A.D., C.N.R.S., Paris, 85 p.

John LYONS (1963) : *Structural Semantics. An Analysis of Part of the Vocabulary of Plato,* Oxford, 1963, 237 p.

John LYONS (1966) : *Towards a "notional" theory of the "parts of speech",* in *Journal of Linguistics,* Vol. 2, no.2, October 1966. (Pre-publication version, kindly communicated by Professor Lyons)

M. E. MARON (1961) : *Automatic Indexing : An Experimental Inquiry,* in *Journal of the Association for Computing Machinery,* Vol. 8, no. 3, July 1961.

Eugene A. NIDA (1964) : *Toward a Science of Translating,* Leiden, 1964, 331 p.

John O' CONNOR (1965) : *Automatic Subject Recognition in Scientific Papers : An Empirical Study,* in *Journal of the Association for Computing Machinery,* Vol. 12, no. 4, October 1965, pp. 490-515.

S.A.D. (Section d'Automatique Documentaire, C.N.R.S., Paris and Marseilles) : most of the publications of the S.A.D. not listed here are listed in MT 1-2, 1965, pp. 21-24.

Gerard SALTON (1966) : *Automatic Phrase Matching,* in David G. HAYS, ed., *Readings in Automatic Language Processing,* New York, 1966, pp. 169-188.

Samuel SELBY, Leonard SWEET (1963) : *Sets, Relations, Functions, — An Introduction,* New York, 1963.

J. P. VINAY, J. DARBELNET (1958) : *Stylistique comparée du français et de l'anglais, Méthode de traduction,* Paris, 1958, 331 p.

Uriel WEINREICH (1963) : *On the Semantic Structure of Language,* in Joseph H. GREENBERG, ed., *Universals of Language,* Cambridge, Mass., 1963, p. 142.

NOTES.

1. This is part of a study supported by the Belgian Scientific Research Foundation (F.N.R.S., Brussels). I wish to thank my supervisors at the University of Liège Professors L. DELATTE, Ph. DEVAUX, M. LINSMAN, and Ir. SIMON, as well as E. SAUVENIER (Maître de Conférences at the University of Liège), J.C. GARDIN (Directeur d'Etudes à l'Ecole Pratique des Hautes Etudes), and the team of the Section d'Automatique Documentaire (S.A.D.) of the C.N.R.S. (Marseilles). None of these persons necessarily agree with everything I say in this paper. As far as possible, I shall indicate the passages which propose views similar or identical to those first propounded in the papers of the S.A.D., which are the main source of this paper as regards the specific field of automatic indexing.

2. This approach is not the only one. In fact, the first studies in automatic indexing have proposed to consider automatic indexing as the extraction by a computer of "significant" words (i.e. words best representing the subjects of the documents) from the scientific texts according to statistical criteria (see M.E. Maron, 1961). But, as M. Coyaud (1964 : 6) has pointed out, the results of this approach still have to be evaluated as to their semantic adequacy, and this takes us back to the problems discussed in this paper.

3. The programming problems are not discussed here. A programme which was written for an early stage of this study and which the computer to index a text according to the cooccurrence of labels obtained by word-for-word "translation" has been written by R. Hoffmann (1966).

4. cf. the set – theoretical distinction given below (ch. III) between the "tabulation-method" and "defining-property method" to describe sets.

5. Example adapted from G.N. Leech (1967).

6. Adapted from both the French and the English definitions given for this subject by F. Lévy.

7. The terms which J.P. Vinay (1958 : 87) calls "termes parallèles" in his theory of translation.

8. See reference EUR 409. f (1964 : 12).

9. In a recent development of SYNTOL, R 1 also serves to relate "predicators" (Q) to the other categories (see A. Borillo et al., 1966).

10. Cp. J. Lyons's recent discussion, within the framework of generative grammar (J. Lyons, 1966).

11. "développement" : see A. Borillo et al. (1966). Note the similarity with the "transitivity" of the "ordering relation" in set-theory : if a R b and b R c, then a R c. Cp. also the "droppint" of Vap : e.g. Violin-merchant derived from violin-selling merchant (Z. Harris, 1965 b : 388) Similar observations in R.B. Lees : e.g. the subject-object relation in auto thief (R. B. Lees, 1960).

12. For the use of such bracketed items see "junction", in the next section.

13. e.g. "Soviet periodical", which is regarded as a specific term with reference to "periodical" in the same way as, say, "rocking-chair" can be considered as a specific term in relation to "chair".

14. "problems ... encountered in" is classified as "S reasoning R1" ("reasoning behind"). This is the only instance I propose to interpret by a "consequence" relation : "problems ... encountered in" (an activity) HENCE "reasoning behind" (this activity) (compare abstract 2 and abstract 29).

15. Another use of "shunting", which will not be discussed here because it goes beyond the purpose of this paper, is the possibility of programming a computer in such a way that it generates from the texts all sorts of 'questions' the user of an automatic indexing system might wish to ask the machine : e.g. to generate a 'question' "aeronautical documentation in the NATO countries" (with dropping of "documentation") from "aeronautical documentation in the NATO countries" (Abstract 31).

16. The last pages of this paper were written at the Department of Linguistics of the University of Victoria, B.C. where I am now working under sponsorship of the F.N.R.S. and of the Canada Council. I am thankful to Professors J.P. Vinay and G. Rondeau for the help they kindly gave me, at their seminars on computational linguistics and in private discussions about this paper.

17. The problems of alternative semantic and syntactic notations – in particular the batteries of tests needed to compute the meaning or syntactic status of ambiguous items – lie beyond the scope of this enquiry.

18. On this point, however, this study may undergo important revisions in future months, as Professor Guy Rondeau has kindly offered to make his automatic parser of English available to the University of Victoria; this parser is the Kuno-Woods algorithm developed at Harvard University and programmed at the University of Montreal. Once this program has been adapted to the needs of this study, it will be possible to envisage a more systematic approach to semantic - syntactic cross-classification than the one proposed here : thus, if the parser assigns the structures (AB)C and A(BC) to both "Swiss libraries, archives" and "historical development, holdings", the valid structure in each case could be detected on the basis of semantic rules (see the role of "semantic networks" in A. Borillo et al., 1966). In many cases this seems to be the only way of discarding the nonsensical syntactic descriptions which all known parsers of English produce next to semantically acceptable descriptions : e.g. "the initial reasoning and thinking" (Abstr. 2) interpreted as "the letter reasoning and thinking"; "data" as the subject in "data processing", or "Lists service bureaus" as "Lists serve bureaus" (Abstract 30).

APPENDIX ONE.

2 ACKOFF, R.L.
Report on a study of scientific communications for the National Science Foundation.
In Shera, J.H., Kent, A., and Perry, J.W., ed. Information systems in documentation, p. 66-74.
New York, Interscience, 1957.

This reviews the initial reasoning and thinking behind the design and construction of a project supported by the National Science Foundation, carried out by the Operations Research Group at the Case Institute of Technology to determine ways "to affect the behavior of scientific institutions that are responsible for the dissemination of recorded information in order to achieve a net improvement in scientific productivity."

5 ADKINSON, B.W.
The federal government and U.S. scientific information.
J. Chem. Doc. 2 (1) : 48 - 50. Jan. 1962.

In addition to presenting a general review of the scope of the Federal Government's role in research and development and a review of the character of the Federal Government's scientific information activities, this article presents a review of the sixpoint program which has been developed by the National Science Foundation for improving the dissemination of scientific information, particularly among U.S. scientists and engineers. The author concludes that the dissemination of scientific information must be recognized as an integral part of research and development.

6 ADKINSON, B.W.
United States scientific and technical information services.
Spec. Libr. 49 (9) : 407 - 414. Nov. 1958.

This discussion of United States scientific and technical information services is divided into the following parts : the present situation in the United States; some of the reasons behind the current problems; the principal scientific and technical information activities of the Federal Government; and a proposed program to remedy this country's scientific and technical information ills.

## APPENDIX 2

42 BAGDASARIAN, S.
Rasshirit' ob' 'em meditsinskoi informatsi Plea for the expansion of medical information.

Med. Rabotnik (Moscow) 53 : 3. July 2, 1957.

The ever increasing need for a more efficient organization of medical information in the Soviet Union indicates that the present decentralized and inefficient system of medical documentation should be replaced with a centralized system. The centralization of medical documentation would permit a better utilization of means and personnel already available to various organizations engaged in medical documentation and better dissemination of information through improvement in methods of presentation and enlargement of the scope of medical documentation. ADI

47 BALKE, S.
Wirtschaft und Dokumentation Documentation and economy.
Nachr. Dok. 12 (1) : 1 - 5. Mar. 1961.

Fundamental problems of documentation are discussed from the professional and organizational point of view. It is stated emphatically that the expert must be provided with the material for his work whatever the financial cost. Development in documentation in West Germany must take the form of cooperation between institutions staffed by highly qualified personnel. Complete centralization as in the Soviet Union is neither possible nor desirable. Documentary activity will be coordinated at national level by the proposed Institut für Dokumentationswesen. ADI

7 ADRESSENLISTE de Dokumentationsstellen der DDR List of documentation centers in the German Democratic Republic.
Dokumentation (Leipzig) 4 (1) : 12 - 17. Jan. 1957.

The address is given for each of 82 documentation centers. Additional information concerns availability of microfilm and photocopying services or photocopying service only, the subject of primary interest to the center, and coverage provided over pertinent Soviet periodical literature. ADI

25 ARCHIVE, Bibliotheken und Dokumentationsstellen der Schweiz. Führer durch die Schweiz. Dokumentation Archives, libraries and documentation centers of Switzerland. Guide through Swiss documentation. Ed. 3. Bern, Sekretariat der Schweiz. Vere in für Dokumentation, 1958. 144 p.

An annotated index to 459 of the most important Swiss libraries, archives, documentation centers, etc., arranged by geographical location. Descriptive data include historical development, holdings, special collections, classification systems used, and user information. The annotations are written in the language used in the pertinent geographical region (German, French, Italian). Alphabetical indexes of institutions, subjects, and personnel are included.
ADI.

29 ASMONAS, V.
Systems of scientific and technical information services : long range planning.
Rev. Doc. 27 (2) : 81 - 85. May 1960.

A proposal for the establishment of an international technical information system which would be based on three main international centers (one for the Americas; one for Europe, the Near East and Africa; one for Asia and Australasia). In addition, there would be a national center for each country, and a center for each scientific or technical discipline. The paper also includes discussion of problems which would be encountered in the development of such a system.

30 ASSOCIATION OF DATA PROCESSING SERVICE ORGANIZATIONS.
Directory of data processing service centers, 1961.
Abington, Pa., 1961, 32 p.

Lists service bureaus in Canada and the United States by province or state, and by city within province or state.

31 ASSOCIATION OF SPECIAL LIBRARIES AND INFORMATION BUREAUS.
Aslib Aero Group Conference, 1957.
Aslib Proc. 9 (8) : 246 - 250. Aug. 1957.

Summary of symposium on aeronautical documentation in the several NATO countries, including description of efforts in Canada, France, Holland, Federal Republic of Germany, U.S.A., and Italy.

38 AVRAMESCU, A., and CANDEA, V.
Introducere in documentarea stintifica Introduction to scientific documentation.
Bucuresti, Akademia, R.P.R., 1960, 519 p. Ref.

The book comprises four chapters and a subject index. The first chapter lays stress on the role and significance of documentation. The authors recount the origins and development of documentation and its present organization abroad and in Rumania, especially in science and technical subjects. An attempt is made to clarify some conceptions, make the terminology more precise and to explain the difficulties connected with documentation. In the second chapter the authors give us the main sources of documentation. They discuss card-indexes, libraries and institutions dealing with documentation. The third chapter points out the different kinds of documentation according to the source, content and value of their information. These kinds are : graphical, pictorial, auditory, visual, etc... The fourth chapter shows the stages of documentation including research work, planning of documentation, collecting and investigating the sources and use of documents. There is a very useful supplement of bibliographic terms in Rumanian, Hungarian, Polish, Serbo-Croatian and Spanish. Unesco Bibliog., Doc., Terminol. 1 (2) : 32-33. May 1961.