Exploitation of Abstracts by Applying Machine Translation Techniques

J. W. PERRY*

Center for Documentation and Communication Research,
Western Reserve University, Cleveland, Ohio

I. SUMMARY AND INTRODUCTION

During recent years, methods have been developed which permit abstracts relating to factual information, particularly in the field of science and technology, to be encoded and thus made amenable to searching, selecting, and correlating operations performed by appropriately designed automatic electronic equipment (1). Extensive tests in the field of metallurgy have demonstrated the practical utility of such methods for providing information services essential to the efficient planning and conducting of research and development programs (2). These tests in metallurgy have been conducted by encoding several years of current metallurgical literature. Such encoding of current literature has been coordinated with its abstracting, alphabetized indexing, and coding for hand-sorted punched cards. The economies in processing costs so attained give promise of being of greatest practical importance (3). Additional economies have also been achieved by using automation techniques to accomplish some of the operations involved in producing encoded abstracts for machine searching. Specifically, the machine translation technique that is sometimes called "automatic dictionary look-up" has been applied to replacing the terms of natural language in grammatically standardized abstracts by semantic codes that denote the principal aspects of meaning of each specific term (4). Such encoding makes generic concepts readily available for defining the scope of searches to be performed automatically. At the present time, the grammatically standardized abstracts make it possible, when defining the scope of search, to take into account those relationships that are determined either observationally or experimentally and are expressed by appropriate phrasing or sentence structuring when writing scientific and technical papers (5) The grammatically standardized abstracts also provide orderly arrays of terms for semantic encoding as outlined above.

* Present address: College of Engineering, University of Arizona, Tucson, Arizona.

These methods, particularly when coordinated with other methods including traditional abstracting and indexing, permit current literature to be processed at low cost to provide information services of unusual versatility and capabilities (2). The question arises "What shall or can we do about the backlog, that is to say, the scientific and technological information of past years?"

To ensure the best results attainable at the present time, it would appear preferable to reprocess the backlog of scientific and technical literature by the same general type of methods now being applied to the world's metallurgical literature. Specific methods would have to be worked out for each field as has been done for metallurgy (1, 4, 5). But regardless of how such methods are worked out in detail, their application to the backlog would doubtless attain the greatest possible effectiveness if the entire backlog were scrutinized by experts in various subject fields so that these experts could make decisions as to what features of information in each backlog publication are of sufficient importance to warrant recording for machine literature searching purposes. In this way, grammatically standardized abstracts would be generated for encoding by automatic techniques already in operational use. If the attempt were made in this way to generate grammatically standardized abstracts of the backlog literature in science and technology, the amount of effort and corresponding costs in terms of time and money might be excessively large. Considered from a slightly different point of view, the re-examination of the backlog by subject experts may be regarded as a theoretical possibility that can scarcely be realized because of practical problems of personnel and costs.

The question arises as to whether such re-examination could not—in large measure at least—be avoided for those papers for which abstracts have been previously prepared. Such abstracts represent the results of much effort on the part of subject experts to state in concise fashion the important information elements in the original papers Such information elements are expressed in previously prepared abstracts by means of phrases and sentences in natural language (e.g. English, German, French, Russian). In encoded abstracts essentially the same information elements must be expressed in grammatically standardized form. It follows that previously prepared abstracts could be exploited for machine searching if such abstracts could be submitted to a translating process which would convert the phrases and sentences of natural language into the grammatically standardized forms of expression used in the encoded abstracts. It is with this question and possible approaches to its solution that this paper is concerned.

II. GENERAL STATEMENT—ABSTRACTS IN RELATION TO INDEXING, CLASSIFYING AND ENCODING

In the field of science and technology, abstracts have, in the past, served a number of important purposes. Perhaps the most important has been to enable a minimum of reading effort to yield a maximum return in understanding the principal conclusions and related facts reported in full length papers. Abstracts have also facilitated the generation of extensive subject indexes whose purpose has been the identification of those abstracts and papers that may be of pertinent interest to a given research and development problem.

The number of abstracts necessary to report publications in a given field, for example chemistry, has increased rapidly in recent decades with the continuing expansion in research and development budgets. The corresponding increase in the bulk of indexes has made their use more and more time consuming.

Studies devoted to evaluating the capabilities and limitations of subject indexes have led to the conclusion that they are inadequate in the form in which they have been previously produced to efficient servicing of many important information requirements, especially those whose scope is such that one or more generic concepts must be specified (6). These inadequacies in subject indexes and a mounting concern to maintain or, preferably, to increase the level of efficiency in scientific research and technical development have provided the motivation for applying and developing various automatic and semi-automatic devices to accomplish the identification of abstracts and corresponding papers that are of pertinent interest to a given research or development problem. The evolution of machine literature searching during the past fifteen years has resulted in the development of practical systems in which a series of index entries or even entire abstracts are encoded for machine selection (1).

It is important to note the close relationship between abstracting, encoding for machine searching, the traditional methods of indexing, and classifying in the conventional sense of arranging into groups patents, documents, or other items.

These four processes, namely, abstracting, indexing, classifying, and encoding for machine searching, could be performed independently by using original scientific and technical papers as the starting point for each of the four different operations. Analysis of the four processes reveals that they all involve the three following preliminary steps: (1) reading and understanding the paper being processed; (2) Deciding what aspects of subject content are important; (3) Selection of terminology or other symbolisms to designate the important aspects of subject content. As shown in Fig. 1, these three basic preliminary operations are followed by different subsequent steps, in particular the composing of sentences in writing literary style abstracts, the establishment of index entries and their subsequent alphabetizing or otherwise arraying in the case of subject indexes, the assignment of class designations with classification systems, and the production of standardized (telegraphic) abstracts for subsequent encoding preparatory to machine searching.

It is true, of course, that there are considerable differences of a structural nature between abstracts, subject indexes, classification headings, and encoded abstracts for machine searching. These dif-

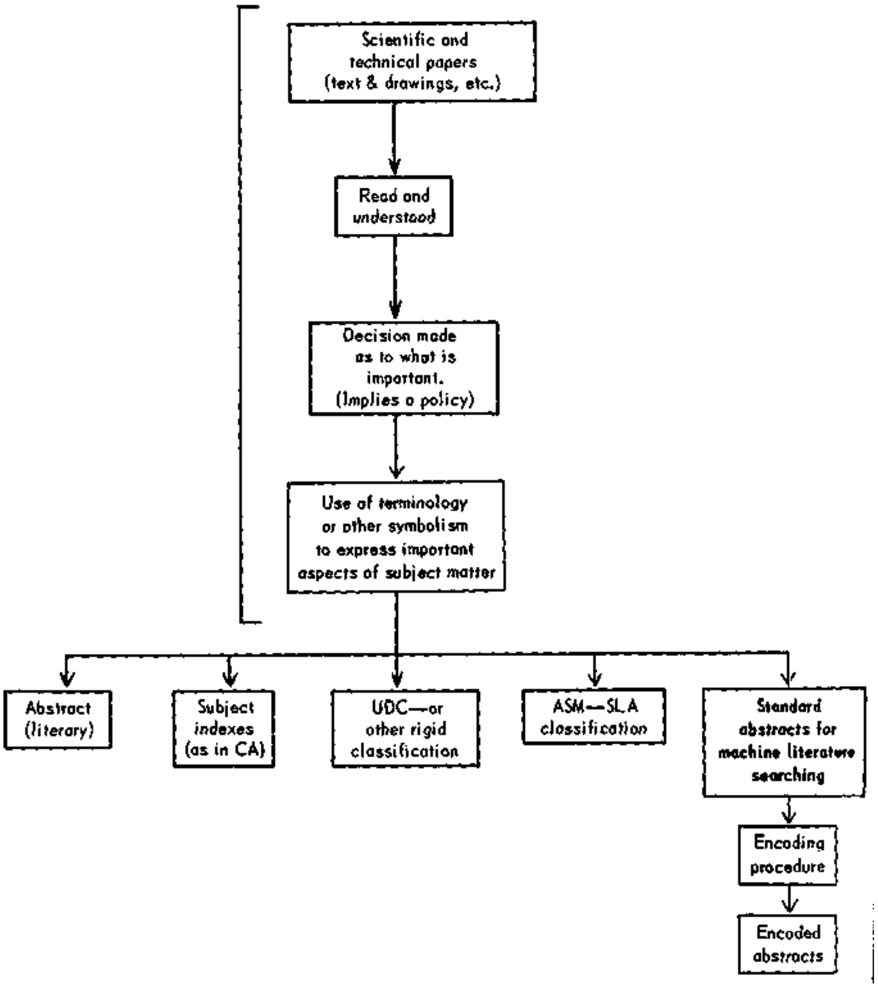
Scientific and technical papers (text & drawings, etc.)
Read and understood
Decision made as to what is important. (Implies a policy)
Use of terminology or other symbolism to express important aspects of subject matter
Abstract (literary)
Subject indexes (as in CA)
UDC—or other rigid classification
ASM—SLA classification
Standard abstracts for machine literature searching
Encoding procedure
Encoded abstracts

Fig. 1. Basic Operations Common to Generating Literary Abstracts for Machine Subject Index, Classifications and Encoded Abstracts for Machine Searching.

ferences mean that the second and third of our three preliminary basic operations would be, and indeed are, conducted somewhat differently, depending on whether the eventual purpose is to produce an abstract, subject index, a classification, or encoded abstract. Nevertheless, close similarity in character of the four underlying processes have a number of important consequences both practical and philosophical. Perhaps the most fundamental consequence is the fact that the overall result of the three basic steps is to conduct an analysis of the subject contents of a given paper in terms of those concepts, basic principles and scientific theories that are currently accepted as valid. This relating of the subject contents of a paper to a broader background of scientific principles and theory involves, from one point of view, an increasing measure of abstraction over and above that which accompanies the interpretation of observations to phenomena in terms of previously developed scientific and technical concepts when writing reports, papers, patents, etc. Another consequence is a certain elimination of individual facts and corresponding reduction in amount of recorded detail.

This situation may be expressed in terms of information theory by saying that there is a certain loss of information. It is important to note that the degree of such loss of information varies greatly depending on whether we are generating an abstract, subject index entries, classification headings, or encoded abstracts for machine parching. Depending on policy, the relative loss of information for each of these four types of output may vary within wide limits. At the same time it is true, as a general rule at least, that the relative loss of information is at a minimum with the literary style abstracts, particularly when these are written to be informative in character. The loss of information is, in general, very little, if any, greater with the encoded abstracts for machine searching. Indeed such encoded abstracts may contain much more information than very brief so-called descriptive abstracts. The greatest loss in information occurs in assigning class designations which, as a rule, do little more than indicate that the subject contents of a given paper pertain to one or more classification subdivisions whose scope is usually defined by a single generic term or by some conjunction of generic terms. The totality of index entries pertaining to the subject contents of a given paper usually is more informative than a classification heading. On the other hand, taken separately, each individual index entry, as encountered in alphabetized or similarly arranged lists, usually pertains to a single narrow feature of the subject contents of the original document. It is particularly important to note in this connection that the degree to which information concerning the subject contents of the original document is recorded can be made substantially equal for both literary style abstracts and also standardized abstracts for encoding for machine searching.

In formulating information requirements, that is to say, in designating a particular range of information that may be of interest to a given research or development problem, both narrow-range specific terminology (e.g. the names of individual substances) or generic terminology (e.g., the broad range of different specific kinds of a general process, such as deterioration) may be important. It is equally likely that a specific term may relate to a certain kind of process (e.g. Hohenstein arc-welding) and a generic feature of a search may relate to a broad class of substances (e.g. alloys containing titanium as principal element).

The analysis of the meaning of individual terms makes it possible to assign to terminology (in particular, terminology of specific character) corresponding codes which indicate how a given term relates to one or several generic terms. This means that it is relatively easy—once semantic codes of the indicated type have been constructed—to convert the recording of specific terms into codes which make it practical and convenient to direct searches to any one generic concept or to their logically defined combinations. At the same time, definition of the scope of a search may make use of appropriate specific terms whose complete codes designate them uniquely and unambiguously.

III. TYPES OF TERMINOLOGY USED IN ABSTRACTING, INDEXING, CLASSIFYING, AND ENCODING

With these facts in mind, it is instructive to consider the kind of terminology that is used to record the varying amounts of information found in abstracts, subject index entries, classification headings, and encoded abstracts for machine searching. In general, we observe that the terminology used in abstracts is somewhat more generic in nature than that found in the original papers, but that the terminology in abstracts is not as generic in character as is frequently involved in specifying the scope of an information requirement.

For the most part, much the same kind of terminology is used in generating subject indexes as is used in writing abstracts with perhaps, at most, no more than a moderate tendency, depending on the field concerned, for subject indexes to be constructed with somewhat less generic terms. This has the consequence that subject indexed are, as a rule, much more useful for identifying papers of pertinent interest to questions and information requirements of narrow scope or character. As a further consequence, subject indexes tend to be less useful for broader searches. The latter are satisfied better by classified arrangements (e.g. of patents or similar documents) provided, of course, that the grouping established by the fixed classification scheme corresponds to that required by the person seeking information. This latter point may be stated somewhat more precisely as follows. Classification headings consist, as a rule, as already noted, of conjunctions of concepts of more generic scope; and if such conjunctions correspond to those which specify an information requirement, then the classification system will provide highly satisfactory results. The difficulty with such systems is that the conjunction of concepts that define an information requirement may be quite different than any one of the limited number of concept conjunctions established when setting up the classification system. The continuing development of new concepts in any field of active research adds an additional element of difficulty to applying conventional classification methods with success. 1

With the possibility of encoding specific terms so as to indicate their relationship to generic concepts, it has become possible to develop encoded abstracts which make available any desired combination or conjunction of generic or specific terminology or both as a basis for defining selecting operations to meet information requirements.

IV. SYNTACTICAL RELATIONSHIPS IN ABSTRACTING, INDEXING, CLASSIFYING, AND ENCODING

Preceding discussion has directed attention to the extent of use and also the mode of use of specific and generic terms in providing literary style abstracts, entries for subject indexes, classification headings, and encoded abstracts for machine literature searching. Both specific and generic terminology pertain to such categories of concepts as substances, devices, processes, properties, functions, conditions, and personalities. In writing reports, papers, etc. to record the results of observations and experiments, relationships between concepts of the above mentioned types are generally expressed by various grammatical or syntactical devices. Considered from the point of view of the extent to which such relationships are recorded, it is perhaps immediately apparent that first place in this respect must be accorded to literary style abstracts. At the other extreme, relationahips of syntactical character are recorded to a minor degree and, in fact, in many cases scarcely at all in the entries of subject indexes or the headings of classification schemes. The extent to which syntactical relationships are recorded in encoded abstracts for mahine literature searching may vary within wide limits. Policy decisions as to procedures in this regard for encoding abstracts can be expected to be controlled principally by the two factors: (1) Effectiveness of recorded syntactical relationships as characteristics for providing useful discriminating power in connection with automatically performed searching operations and (2) the cost of establishing such relationships in a consistent fashion. More specifically, particular attention in this connection must be directed to the cost of recording syntactical relationships in encoded abstracts, and the cost of taking such relationships into account when converting information requirements into searching programs to be performed by automatic equipment. If, on analysis, it is found that commensurate benefits are provided by recording a certain set of syntactical relationships, then appropriate policy decisions are advisable in setting up procedures for generating encoded abstracts. In other words, the extent to which syntactical relationships are recorded in encoded abstracts must be considered from the point of view of benefits vs. costs. In formulating procedures for encoding syntactical relationships, logical considerations can provide guidance in avoiding inconsistencies and ambiguities. It should be emphasized, however, that the mere fact that a given type of relationship is logically valid does not constitute justification of a decision to record relationships of the given type in encoded abstracts for machine searching. It may be concluded, therefore, that in setting up a machine searching system the decision should be to record in the encoded abstracts the simplest set of syntactical relationships that suffice to provide adequate discriminating power. In formulating this practical principle, it is not intended to imply that it is an easy or simple matter to determine in a given set of practical circumstances just what syntactical relationships are needed to provide adequate discriminating power. In fact, at the present time such decisions must be made largely on the basis of experience and judgment. In the near future a firmer more objective basis for making such decisions should result from the mathematical analysis of the performance of mechanized documentation systems.

V. RESUME OF COMPARISON OF ABSTRACTING, INDEXING, CLASSIFYING, AND ENCODING

This brief review of the kind of terminology and the kind of syntactical relationships expressed and recorded by literary abstracts, the entries of subject indexes, the headings of classification systems, and encoded abstracts for machine literature searching may suffice-to make the following points: 1. Not only are there close similarities between literary style abstracts and encoded abstracts for machine literature searching with respect to the amount of information these two types of abstracts contain, but there are additional similarities in the specific and generic concepts and syntactical relationships that are used to express such information. (Considered from the point of view of procedures, abstracts as encoded for machine literature searching might be regarded as literary style abstracts which have undergone severe editing as to phrasing. Furthermore, as a result of the encoding of specific terms, the encoded abstracts make available both specific terms and also their related generic concepts as a means for defining the scope of search to be performed by automatic equipment). 2. Although the encoding of subject indexes—as outlined in Figure 2—appears not only possible but seems virtually certain to provide useful results, nevertheless, the lesser average amount of information contained in subject indexes and in particular the very extensive elimination of important syntactical relationships during the generation of subject indexes means that the encoding of subject indexes, as found for example in “Chemical Abstracts,” will provide far less discriminating capability when submitted to machine searches than would be the case if the literary style of abstracts, as found in “Chemical Abstracts,” were converted to the encoded form. 3. Subject headings of classification schemes, because of their very extensive elimination of important detail by assigning of headings constructed for the most part from generic terminology, will also, if encoded, provide much less discriminating power than would be the case if literary abstracts are used as a starting point for encoding operations. These three conclusions are scarcely a matter of speculation but have been demonstrated by experimental work conducted at the Center for Documentation and Communication Research (7).

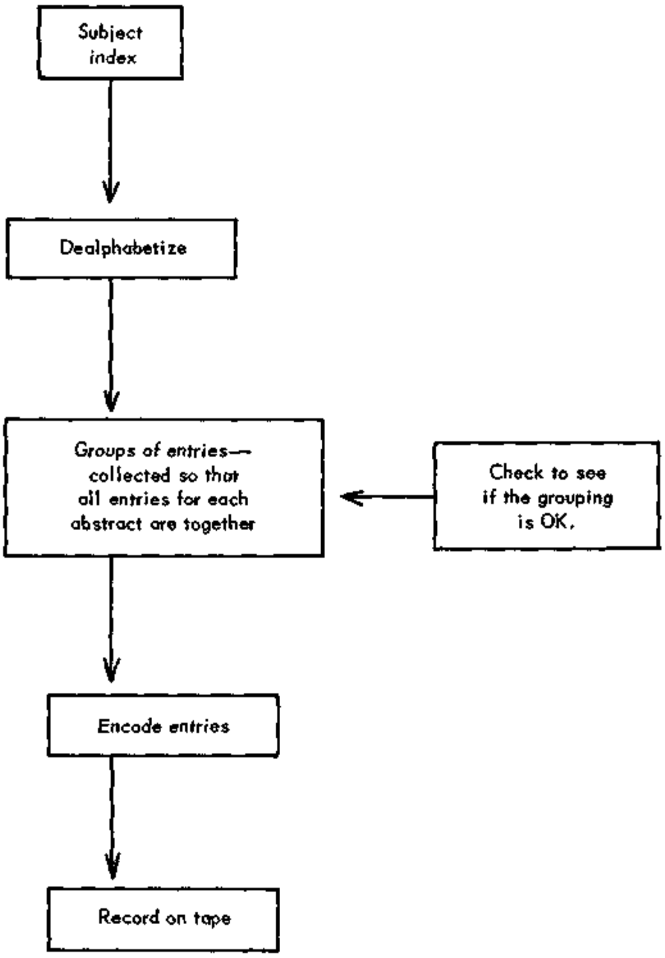
Subject index
Dealphabetize
Groups of entries— collected so that all entries for each abstract are together
Check to see if the grouping is OK.
Encode entries
Record on tape

Fig. 2. Sequence of Steps in Encoding Subject Indexes.

VI. THE ENCODING OF LITERARY STYLE ABSTRACTS

During the period of November, 1955, to January, 1957, when our methods for encoding scientific and technical abstracts in the field of metallurgy were being formulated in detail, several thousand informative abstracts as found in the metallurgical section of "Chemical Abstracts" and "Metallurgical Abstracts" were read by an expert in encoding procedure. The subject contents of the abstracts were expressied with the aid of a standardized set of syntactical relationships (5). This drastic editing operation was conducted by making use of two general types of syntactical devices: 1. The assignment of special codes (role indicators) which indicated the mode of involvement of various terms, most of which were identical with those found in the literary style abstracts; 2. The organization of the building blocks so generated (that is to say, the combinations of role indicators with accompanying terms) into ensembles analogous to phrases, sentences, paragraphs, and complete messages in ordinary literary language. The beginning and ending of phrases within sentences, sentences within paragraphs, paragraphs within messages were denoted by specially assigned symbols which may be regarded as analogous to the punctuation of ordinary literary style abstracts. It is particularly important to note the following: 1. in the standardized abstract the syntactical devices used, though analogous to those of ordinary language, are much smaller in number; 2. as a consequence in the encoded abstracts certain distinctions are not made which are logically valid and which could be and indeed often are made in ordinary literary language. Thus a given syntactical relationship, as specified in the encoded abstracts, may correspond to a considerable variety of alternate ways of expressing either the same relationship or a set of closely related though readily distinguished relationships in the literary language; 3. as a consequence of such standardization of modes of expression of relationships in encoded abstracts, the programming of searches in which syntactical relations are used as factors for achieving required discrimination is greatly simplified in comparison to what would be the case if such standardization of expression of relationships had not been carried out; 4. the simplification of programming of information requirements achieved in this way makes it possible to conduct searching operations of the highest practical utility by means of automatic high-speed equipment whose cost is a tenth or less than that of the so-called general purpose computers. I

It is important to emphasize that during the period when the literary abstracts were being converted by human effort into standardized form (so-called telegraphic style) the terminology of the literary abstracts was subjected, at most, to very minor alteration while their phrasing, that is to say the expression of syntactical relationships, was drastically changed. Once the standardized abstracts had been so generated, an automatic encoding process closely similar to the dictionary look-up methods previously developed for machine translation was applied to replace specific terms by their codes whose elements designate related generic concepts in line with the meaning of the specific terms. I

Experience in encoding well over 10,000 abstracts in metallurgy during the period since November, 1955, has established a number of conclusions. I

1. The task of writing out standardized abstracts to record the subject contents of the literary style abstracts or of a corresponding paper can be formulated as a series of simple rules than can be expressed concisely and that can be learned in a short period of time.

2. These rules have been worked out in such a way that their application in generating standardized telegraphic style abstracts is subject to very little doubt. This has been confirmed by the virtual

absence of differences in ways of expressing relationships when the same literary style abstracts or full-length papers were processed by different persons in an independent fashion to produce standardized telegraphic style abstracts.

3. The set of relationships, as specified by the rules for standardized telegraphic abstracting, provides effective guidance in selecting those terms which are appropriate to recording important features of the subject contents of literary style abstracts or papers that are being processed. In other words, the rules for generating standardized abstracts provide, as it were, a framework of relationships into whose slots appropriate terminology may be fitted, and in addition the rules of telegraphic abstracting provide guidance in selecting the terms for placing in the appropriate slots. Summarizing these considerations, it might be said that the rules for generating standardized telegraphic abstracting not only state policy as to which features of subject contents are to be expressed but also provide guidance in selecting appropriate terminology for expressing such features.

VII. ENCODED ABSTRACTS—A STANDARDIZED ARTIFICIAL LANGUAGE

Considered from a linguistic point of view, the rules for telegraphic abstracting may be said to constitute the grammar of a standardized artificial language. It is perhaps obvious that alterations in this grammar can be made as may be appropriate in establishing policies and procedures to meet different types of information requirements or to adapt this specialized form of artificial language to a new field. It should be noted further that the grammar of such an artificial language, though developed in its present form by English-speaking persons, is directed to recording those relationships that are of particular importance in science and technology whose international character is generally recognized. It is to be hoped that further development of machine language will further accentuate its internationa character.

As already noted, the final step in the encoding of abstracts for machine searching is to replace the individual terms in standardized abstracts by corresponding codes. These previously established codes record the results of analyzing the meaning of words or terms, some of which may consist of two or more words, so as to indicate their relationship to generic ideas. A code dictionary of this type has been worked out for approximately 20,000 terms. The usefulness of this kind of coding in conducting searching and selecting operations by automatic equipment has been repeatedly demonstrated during recent years. The practical effectiveness of such searching and selecting operations provides conclusive evidence that the defined meanings of scientific and technical terms are sufficiently precise and invariable to serve as the basis for analyzing the subject contents of documents on the one hand, and for converting information requirements into machine searching programs on the other hand. As a consequence, automatic equipment can perform matching operations between the encoded characteristics of the subject contents of papers on the one hand and of information requirements on the other hand.

It must be kept in mind, however, that complete precision of definition of the meaning of terminology is an ideal which may be compared to the absolute zero of physics or the completely reversible reaction of chemistry. This ideal is very closely approached with many scientific and technical terms, at least at a given period in the history of science and technology. However, it must also be recognized that the meanings of various terms may undergo change as science advances. An example is the considerable degree to which the term “chemical element” was redefined under the impact of the concept of isotopes. A more or less extensive revision of codes assigned to terminology may thus be required as scientific research and technical development advance. The more rapid such advance may be in a given field of specialization, the greater will be the advisability of maintaining close scrutiny of the suitability of the codes assigned to specific terms. It must be emphasized, however, in this connection, that an "unsuitable" code for a given term does not mean that the term and its corresponding code are unusable in formulating a search program. Rather an “unsuitable” code constitutes no more than an inconvenience in formulating search requirements. If the proportion of such codes is permitted to become larger, the degree of inconvenience increases correspondingly and may require, in the extreme case, so much additional effort as to render such a coding system impractical from an operational point of view. This point becomes more readily understandable when it is considered that the code itself has the purpose of making more convenient the use of sets of related terms, Such terms, even in the absence of a semantic code, could be combined, in principle at least, into sets but this would require the listing of lengthy arrays of terms with the result that the machine programming would be rendered complex by involving logical sums consisting of large numbers of terms.

In a great majority of eases, as already noted, scientific and technical terms each have a meaning which varies to such a slight degree in different contexts to permit a given term's principal aspects of meaning to be designated by a single code. It is true, on the other hand, that there are certain scientific and technical terms whose meaning varies sufficiently with context to warrant the establishment of different codes for the different meanings of the same term. Such terms (or homographs as they are sometimes called) may be exemplified in English by the structural term “cell” or the operational term “polarization”, whose meanings are quite different when applied to phenomena involved in optics and in electrolysis. Experience to date in generating encoded abstracts indicates that the decision as to which code is appropriate for a given homograph in a given abstract can be readily determined by consulting the context. Thus, the word “cell” when used to refer to a component of a biological organism would be encoded quite differently than when it is used to refer to the basic unit of a crystal or to a small chamber in a building. Up to the present, the frequency of occurrence of such homographs in our encoding operations has been so low that it has not appeared appropriate to develop machine routines which would take into account contextual variations in the meaning of terms in order to accomplish automatic selection of the appropriate code for a given homograph. Our experience to date indicates, however, that selection of the appropriate code could be accomplished with a high order of reliability by taking into account contextual terms or more particularly their meanings as expressed in semantic codes.

The preceding discussion has been presented to make the following points:

1. In our present methods for generating encoded abstracts, the techniques of machine translation are already being used to a limited degree, namely, the dictionary look-up for replacing English language terms by their semantic codes.

2. The relationships designated in standardized telegraphic style abstracts are of such nature that the rules for generating such abstracts constitute a special standardized grammar and this suggests the possibility of further application of machine translation techniques to the generation of encoded abstracts.

In this connection it is perhaps well to emphasize that the relationships specified by the grammatical devices of the standardized abstracts are precisely those which are of major importance from a scientific and technical point of view. Such relationships are, of course, precisely those which will be expressed most clearly and unequivocally in various literary style abstracts written in one or another of the natural languages. Thus it can be anticipated, and indeed it has already been observed, that there is a close correspondence between the relationships expressed in literary style abstracts and the relationships expressed in encoded abstracts. This is true not only when the literary language is English, German, French, Russian, or some other language of the Indoeuropean group, but also, as discussions with Japanese scientists have made clear, it is equally true of abstracts written in Japanese. Although the literary devices and syntactical structure of the Japanese language are widely different from those of the Indoeuropean group, nevertheless the international character of science and technology exerts such a strong influence that the relationships most clearly stated in a literary style scientific and technical abstract will be precisely those that we have been taking into account in preparing our standardized telegrapic abstracts that are subsequently encoded for machine searchings.

VIII. MACHINE TRANSLATION AND SCIENTIFIC EXPRESSION

As recently reported Russian experience in the machine translation of mathematical papers has demonstrated, there is a very marked tendency for scientific and technical relationships of grammatical character to be expressed in much the same fashion by different writers (9). Indeed, as Russian experience in machine translation of mathematical papers indicates, it is, at worst, a mild exaggeration to speak of a quasi-standardization of mode of expressing relationships in mathematical writing. This tendency towards standardization of syntactical devices in a given field of specialization was, of course, taken into account and skillfully applied by the Russians in setting up machine translation procedures as diagrammed in Fig. 3. In this diagram, a text in one of the natural languages, an exemplified by a French input text, an English input text, and a Japanese input text, constitutes the input to an appropriately programmed analyzer. As far as the design of the equipment is concerned, the various analyzers for French, English and Japanese in Figure 3 could be identical, but the programming of the equipment for these three languages would differ depending on the language of the input text to be processed. The output from these different analyzes was a set of instructions for generating the Russian output, namely, a Russian translation. This is indicated in our diagram by the box labelled codification data for providing a Russian output. This codification data is in effect a program for generating Russian

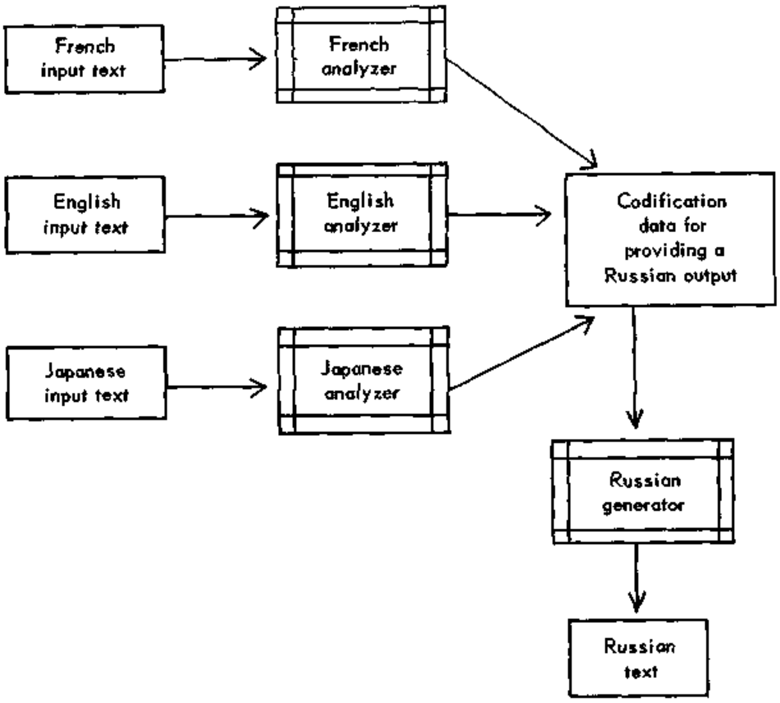

Fig. 3. Coordinated Machine Translation of Input Text in Different Natural Languages

sentences whose building blocks will be Russian terms corresponding to those of the input text and those mode or organization into phrases and sentences is determined by the Russian grammar. It is perhaps obvious that the Russian generator must be regarded as performing two distinct though well-integrated tasks. One of these is to provide the Russian words from the dictionary type memory, and in case of uncertainty as to the appropriate Russian term to select the proper term. The other operation is to organize the words into appropriate phrases and sentences, i.e. to insure correct Russian word order, to supply endings, etc. Since every dictionary, no matter how complete at a given date, must be extended to incorporate new terms, the automatic dictionary which constitutes a basically important unit in the Russian generator will have to be revised and brought up to date from time to time. New or unusual forms of grammatical expressions may confront the portion of the Russian generator which provides endings for words and organizes them into phrases and sentences with a similar kind of problem that the machine cannot handle and that must be brought to the attention of the operator.

IX. MACHINE TRANSLATION APPLIED TO ENCODING OF ABSTRACTS

Figure 4 shows how the Russian mode of conducting machine translation might be incorporated into a general scheme for processing scientific and technical papers to provide various outputs, in particular abstracts for publication, subject indexes, UDC or other classifications, and encoded abstracts for machine literature searching. The bottom row in this diagram indicates how literary abstracts in various natural languages could be first run through an analyzer which has been appropriately programmed for German, English, Russian, French, Japanese, etc. Analogously to the above outlined Russian work in machine translation of mathematical papers, the output from these analyzers might be codification data for providing our encoded abstracts. Next, this output could be run through a machine programmed to generate encoded abstracts with the output being, of course, in a form ready and appropriate for machine searching. The complete parallel between this sequence of procedures and those diagrammed in Fig. 3 is perhaps obvious and, as the discussion in connection with Fig. 3 has already pointed out, newly encountered words would have to be accorded similar treatment when they are to be translated either into Russian or into encoded form for machine literature searching. Stated somewhat differently, this means that words and terms not in the code dictionary would have to be taken into account and entered into the code dictionary in the same way that Russian words to correspond with newly encountered foreign words would have to be provided in the internal memory of the Russian generator when producing output Russian text as diagrammed in Fig. 3.

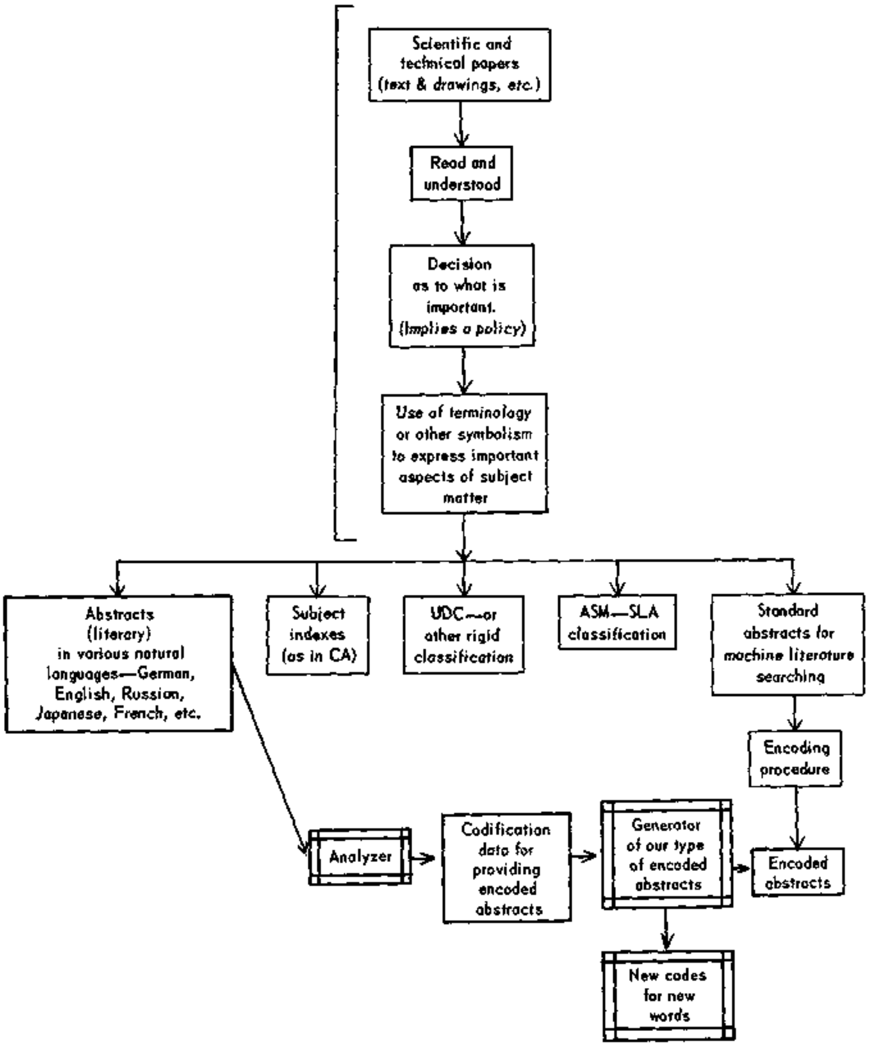
Scientific and technical papers (text & drawings, etc.)
Read and understood
Decision as to what is important. (Implies a policy)
Use of terminology or other symbolism to express important aspects of subject matter
Abstracts (literary) in various natural languages—German, English, Russian, Japanese, French, etc.
Subject indexes (as in CA)
UDC—or other rigid classification
ASM—SLA classification
Standard abstracts for machine literature searching
Encoding procedure
Analyzer
Codification data for providing encoded abstracts
Generator of our type of encoded abstracts
Encoded abstracts
New codes for new words

Fig. 4. Coordination of Processing of Current Publications with Automated Processing of Literary Abstracts into Encoded Form.

It should be noted in this connection that this mode of conducting the encoding of abstracts, as outlined in Fig. 4, would probably be best worked out so that a continuing running check might be made as to the observed contextual relationships between words. Let us assume, for example, that the word "polarization" had been noted in the code dictionary as relating to optics. A check as to the actual occurrence of such a relationship could be made automatically and in this connection semantic factors for optics and the like might prove rather useful and convenient. If the word “polarization” were then encountered in another widely different context, for example, in connection with electrolytic processes, such deviation in the character of the observed context could be detected automatically and brought to the attention of a person who checks and takes appropriate action for such instances.

Before leaving discussion of Fig. 4, it should be pointed out that this kind of application of machine translation techniques to the encoding of literary abstracts would offer the possibility, once the preliminary development of equipment and machine programs had been worked out, of encoding abstracts for machine literature searching at very moderate cost. In previously generated literary style abstracts there are recorded, in effect, the results of very extensive investments of time, energy, and money. The input to the machine translating procedures, as outlined above, could be accomplished at the present time by conventional keyboard operations or in the near futture by character recognition devices when they become available commercially. In this way the step of manual keyboard processing could be obviated.

It should also be emphasized that the machine translation of literary style abstracts into encoded abstracts is to be sharply distinguished from processes sometimes referred to as auto-abstracting whose purpose is to take full length papers and convert them by automatic procedures into abstracts. In auto-abstracting, the machine is called upon to make decisions as to what is important; and auto-abstracting has been proposed as a task to be accomplished with the aid of statistical evaluation by the machine of frequently occurring phrases or sentences. In this way, the machine makes its decision as to which phrases and sentences are most likely to be important, in contrast to auto-abstracting; the abstracts to be encoded, as outlined in Fig. 4, by machine translation methods have been prepared by human experts. In the above indicated conversion of literary abstracts into encoded abstracts, the entire subject contents of abstracts previously prepared by human experts would be processed so as to designate, in a standardized form, those relationships that are important from a scientific-technical point of view and that are useful as a basis for converting information requirements into machine programs. Aside from such standardization and the possible elimination of excessive redundant verbage, the entire information content of the literary abstract would be converted into encoded form. As already noted, such coding would make it possible to use low cost searching and selecting equipment to perform literature searches. The range of diversity of such searches may be made very broad by exploiting generic terminology built into semantic codes or by exploiting the "near-miss" capabilities of appropriately designed equipment or by using automatic correlation techniques. Varying degrees and types of correlation of scattered information can be automatically accomplished.

XI. CONCLUSION

The procedures outlined in this paper, when once developed, should make it possible to convert at low cost extensive files of abstracts, as published for example in "Chemical Abstracts," "Chemisches Zentralblatt," the "Referativnye Zhurnaly" or other abstract periodicals, into a form which will make them readily searchable by equipment at moderate costs. It appears unnecessary to emphasize the practical usefulness of such procedures for converting the results of previous abstracting efforts into a form which can be readily and widely exploited by automatic searching equipment.

APPENDIX

Notes on Development of Machine Translation Programs

From previous discussion it is perhaps obvious that a large portion, in fact almost certainly a major portion, of the development of methods for applying automation techniques to convert literary style abstracts into encoded abstracts for machine searching has been already accomplished. In applying the present semantic code dictionary to fields outside of metallurgy, additional terms would perhaps have to be coded; but their number seems likely to be much smaller for other fields of science and technology than might be anticipated. Consequently, the principal development effort would be directed to establishing (1) procedures for analyzing literary style abstracts in such a way an to produce the codification data for generating encoded abstracts and (2) procedures for using such codification data to generate encoded abstracts.

As already noted, the input, that is to say the literary style abstracts, would be submitted to automatic machine analysis to provide the data needed for generating standardized encoded abstracts for machine searching. This program would, of course, be designed to take into account (1) the grammatical peculiarities of the input language, (2) data required for producing the encoded abstracts, and (3) the operational characteristics of the equipment to be used both during the step of input analysis and subsequent step of generation of standardized encoded abstracts. Since, in principle at least, any so-called general purpose computer having the characteristics of a so-called near-Turing machine can accomplish any logically defined routine, the preliminary development of the two machine programs would be possible in principle, as soon as each step in the processing has been clearly and unambiguously defined.

In working out the input analysis program, the grammar of the input language and the way in which such grammar has been formulated and exploited by others in accomplishing machine translation would be taken into account. As already mentioned, advantages would appear easily attainable by taking account of the stylistic characteristics of scientific and technical languages. More specifically, a principal anticipated advantage is the possibility of developing simpler machine programs whose special design would also guarantee high levels of reliability in conducting input analysis.

In working toward this dual goal of simpler machine programs and high reliability of input analysis, particular attention should be directed not only to the stylistic character of scientific and technical writing and to well-known grammatical features of a given language but also to those types of phrasing which, in effect, constitute unwritten rules that nevertheless are habitually obeyed when composing phrases and sentences in natural languages (8). An example will make this point clearer. Let us consider the following English adjectives: French, pretty, shaggy, old, paper, green, wooden, steel (heere it will be noted that both paper and steel may be used either as adjectives or as nouns in English). Let us observe how two or more of these adjectives may be used as joint modifiers of the English nouns—poodle, girl, bag, bridge. In particular, attention will be directed to the order in which two or more of these adjectives will be used in constructing adjective-noun phrases. Thus, the phrase "shaggy French poodle" is in accord with common use, while interchanging the adjectives as in the phrase "French shaggy poodle" is a word order which is instinctively avoided. "Shaggy old poodle" will be preferred to "old shaggy poodle" and "shaggy old French poodle" will be preferred to phrases in which the three adjectives are arranged in any other order. Note further the similarity in adjective order in the phrases "old wooden bridge" and "new steel spring" which are preferred to phrases in which the adjectives are in different sequence. Examples of this sort can be created in great number. For example, "old paper bag" is certainly preferred to "paper old bag" and "beautiful young French girl" to "young French beautiful girl" or "French beautiful young girl."

It would appear that in English there is a rather definite preference constituting in effect a rule of grammar in the order in which multiple adjectives will be cited in sequence when they all modify the same noun. Further investigations are likely to reveal that the adjective which stands nearest to the noun is the one which designates material from which it is constructed—some feature which is thought of as being perhaps more definitive in the sense that French is more definitive than shaggy or beautiful or young. On the other hand, adjectives which indicate age, e.g. young, old, appear to be regarded as more superficially descriptive than words which indicate that a non refers to an object that has undergone some process. Thus we would have rough welded joint rather than welded rough joint.

In formulating and defining such crypto-grammatical rules— as Whorf calls them—it seems likely that semantic factors may be highly useful and furthermore a semantic factor type of analysis of the meaning of words may turn out to be the key to exploiting this type of grammar rule for machine translation purposes. These purposes, it might be noticed, should not be thought of as being limited entirely to conducting the analysis of input material. It seems worthwhile to undertake formulation of rules than may be applied to check and to confirm the analysis and the expression of relationships by assignment of role indicators. Thus, in conducting the analysis of the phrase “the production of the paper bag” by considering the order of the word “paper bag” as contrasted with the different meaning of “bag paper,” we see that the production of “paper bag” will indicate “bag” as material produced and “paper” as the component; whereas a slightly different arrangement of words as in the production of “bag paper,” the material produced is clearly “paper” and the word “bag” indicates a kind of object for which such “paper” may be used.

It is virtually certain that establishing machine programs for the input analysis of English abstracts will involve the discovery and formulation of many such rules of phrasing and that in the formulation of these rules—in contrast to an assumption often made in describing the development of machine translation—an interaction between the meaning of the words and the grammatical phrasing will be detected, formulated, and exploited.

In English, such development of programs must devote particularly careful attention to word order, or to express the same idea in a slightly different form, in English, word order is, in many phrases and sentences at least, the principal source of clues as to the relationship between words. In the heavily inflected languages, of which Russian is an example, the situation is somewhat different. It is true that the order of words in a Russian sentence is—at least as far as scientific and technical writing is concerned—controlled by rules to a greater extent than may be commonly realized. Thus, for “my black book2 in Russian, one would say моя черная книга, but scarcely черная моя книга which would correspond to “black my book” in English. Similar kinds of rules relating to the order of words and phrases, and of phrases and sentences, appear to be similar in nature in both Russian and English—and in some cases even the rule itself may be closely similar. In Russian we have in addition a much more extensive system of inflectional endings than in English. It appears virtually certain that Russian grammatical constructions in scientific and technical writings are characterized by a much higher degree of redundancy than is the case with English. In considering these points, it must not be overlooked on the other hand, that deviations from standard word order—or the more frequently encountered word order, to be more precise—may be encountered in both languages and such deviations often serve the purpose of providing emphasis. Consequently, a considerable amount of caution is advisable before arriving at the conclusion that only a certain specified word order can be expected to be encountered and that, accordingly, it is to be regarded as a true or probably true standard.

In establishing various word order rules for English and also for any other language, the statistical approach would appear advisable. That is to say, the most practically useful program of analysis must be worked out to cope with an acceptably high percentage of the sentences and constructions encountered without imposing the requirement that the program shall be able to analyze successfully every conceivable sentence that might be constructed in accordance with the rules and practices of grammar. Here actual experience in analyzing a sufficient body of text to be statistically significant would be essential in arriving at programs which are effective in the majority of cases encountered while avoiding the disadvantage of excessive complexity. Unusual sentence construction should be detected automatically by the machine program and thus brought to the attention of the human operator for appropriate handling.

In developing programs for processing literary style abstracts as input, it is to be expected that the purpose of such processing will be decisively important. This can be illustrated by directing attention to the problem of translating the reflexive form of Russian verbs into English. For example, consider the two Russian sentences, Сахар легко растворяется and. Сталь широко применяется. In both of these Russian sentences, the verb is in the third person, singular of the present tense, and is also reflexive as indicated by the last two letters. If our purpose were to translate such Russian sentences into English, the first would be translated as "Sugar dissolves easily in water," whereas the second Russian sentence would be translated as "Steel is widely used." Thus, in the first Russian sentence an intransitive verb is used to translate the Russian reflexive whereas in the second the English passive voice is appropriate. Consequently, for an English output one would have to arrive at a way of distinguishing the translation of the verb in the first Russian sentence from the translation of the verb in the second Russian sentence. On the other hand, considered from the point of view of establishing encoded abstracts, such a distinction might be unnecessary as both сахар " sugar" and сталь "steel" are things acted on. In one case the process is "dissolved" and in the other case it is "used." These two Russian sentences may be regarded as providing an example of how our purpose in providing enoded abstracts as the output may strongly influence the machine program. This means, furthermore, that previous experience in geerating standardized encoded abstracts for machine searching is of essential importance in developing the type of machine translation prcessing with which this paper is concerned.

To consider a further example, let us note that the use of the instrumental in Russian with nouns denoting a period of time such as ночью, "by night," or днем, "by day," would doubtless have to be distinguished from the use of the instrumental to designate a means by which something is accomplished, as in the Russian sentence, натрий легко режется ножом, where ножом is the instrumental of the noun нож, "knife." In this connection one must also keep in mind that the instrumental may be used in such expressions as Этот предмет является ножом in translation "That object is a knife", or literally "That object shows itself as a knife." This emphasizes the obvious point, that the clues to meaning provided by Russian case endings must be considered in conjunction with other clues, either of a grammatical nature (inflectional endings or word order) or of semantic character (features of meanings of words).

An example of a problem which is encountered in German is the compound word, which, in spite of the German printing convention of omitting spaces between the words, has much in common with parallel English expressions. Thus for example one encounters in German such words as Lebensversicherungsgesellschaft, "Life Insurance Co." or Frachtbeförderungsgesellschaft, "Freight Forwarding Agency." It is to be observed that it is not acceptable to say in German, Versicherungslebensgesellschaft any more than it would be acceptable in English to say "Insurance Life Co.", and the rearranged compound word, Beförderungsfrachtgesellschaft, would also be contrary to German usage. In some cases, however, the reversal of the order of words in a compound has the same effect on the meaning in German as in English. Thus Papiersack means "paper sack," i.e. sack made from paper, while Sackpapier means "sack paper." Similarly Holtzbrücke is "wooden bridge" while Brückenholtz is "bridge wood", that is to say, wood for constructing bridges. When one considers the close family relationship between the German and English languages these similarities are, perhaps, not surprising. It will be recalled in the case of German, however, that the word order in constructing sentences often deviates greatly from English so that many word order rules, particularly those which involve the position of verbs in a clause or sentence, are formulated quite differently for German than for English. On the other hand, some English rules might turn out to be more or less directly applicable to German in which one may say junge französische Frau, "young French woman," but scarcely französische junge Frau, corresponding to "French young woman" in English.

When we step outside the Indoeuropean group of languages, we may expect to encounter problems of program formulation of a widely different character. For example, various important categories of words in the Japanese language are difficult to relate to our familiar parts of speech, such as nouns, adjectives, verbs, adverbs. From the Japanese point of view, the situation is quite simple as their words may be grouped into two categories, (1) those that are inflected and (2) those that are not inflected. This latter group includes both the so-called "na" words that name things, conditions, abstract and concrete concepts, basic actions and particles that denote various relationships—rather like prepositions and conjunctions. The inflected words (so-called "working" words) may function like our adjectives by defining, or limiting the "na" words or to indicate action involving the "na" words. As a consequence, an English-speaking person may find it very misleading—even impossible—to make use of the familiar noun, verb and adjective categories in attempting to understand Japanese syntax.

Various particles in Japanese are used with a significance similar to our role indicators. For example, the particle, wo for the most part designates that which is the object of an action. Other particles in Japanese indicate relationships which are virtually impossible to translate directly and understandable only in terms of a relationship which they designate between other words in a sentence. Native Japanese speakers have no more difficulty in using such particles than we have in deciding that "old paper bag" is preferable to "paper old bag." Such connective particles may appear to be concerned with parts of speech akin to our role indicators but with the significance of the particles strongly influenced by the words with which they are used; and here it would appear probably that analysis would reveal a relationship between the meaning as determined by the phrasing which would be more or less closely analagous to the rules which determine the order of two or more adjectives used with a single noun in English. In conducting such an analysis in Japanese, it seems likely that semantic-factor or at-the-source type of analysis would not only provide useful clues to detecting and formulating such syntactical rules but also to exploiting them in developing machine programs. In this connection striking similarity between certain Japanese ideographs and our semantic factors is worthy of particular mention. Thus, the Japanese character, gaku, has much the same significance as our semantic factor, S-CN, meaning science, while another ideograph designated as den in combining forms appears in combinations of characters which indicate such ideas as ammeter, electric light, volt, electric welding, etc. The similarity with our semantic factor L-CT for electricity, is truly astonishing. Such a striking degree of correspondence is not observed of course, between all our semantic factors and the Japanese ideographs. Nor is such correspondence to be expected. Obviously our analysis of terminology has been conducted for certain specialized purposes and consequently has not followed the same path as the development of words and ideas in Japanese. Nevertheless, at least in scientific and technical terminology, the degree of correspondence may prove sufficiently extensive to provide considerable practical aid to developing a Japanese analyser for abstracts written in the Japanese language.

REFERENCES

1. Perry, J. W., Kent, Allen, and Melton, John L., "Tools for Machine Literature Searching," Interscience, New York, 1958.
2. Rees, Janet, and Kent, Allen, Mechanized Searching Experiments Using the WRU Searching Selector, "American Documentation," (1958), pp. 277-303; Perry, J. W., Kent, Allen and Melton, John L., op. cit. Chapter 15, "Introduction to Analysis of Questions" by Jessica Melton and J. W. Perry.
3. Rees, Alan M., "Coordination of Procedures for Abstracting, Indexing and Encoding of Metallurgical Literature." Ph.D. Thesis (in preparation). See also Perry, J. W., Kent, Allen, and Melton, John L., op. cit. Chapter 8, "Telegraphic Abstracts—Coordination with Other Information Processing" by Alan M. Rees, Allen Kent and J. W. Perry.
4. Perry, J. W., Kent, Allen, and Melton, John L., op. cit. Chapter 9, "The Semantic Code" by John L. Melton.
5. Perry, J. W., Kent, Allen, and Melton, John L., op. cit. Chapter 5, "Procedures for Preparation of Abstracts for Encoding" by Jessica Melton.

6. See, for example, Bailey, M. F., A History of Patent Office Classification*, “Journal Patent Office Society*.” *28* (1946), pp. 463-507, 537-575.
7. Perry, J. W., Kent, Allen, and Melton, John L., op. cit. For a more extensive state of the methods of mechanical translation, see Booth, Andrew D., Brandwood, L., and Cleave, J. P., “*Mechanical Resolution of Linguistic Problems*,” Academic Press, New York, 1958.
8. Whorff, Benjamin Lee (John B. Carroll, editor), “*Language, Thought and Reality*,” John Wiley and Sons, New York, 1956, cf. especially p. 93.
9. Andreyev, N. D., Chapter 49, this volume.