[Proceedings of the *International Conference on Scientific Information*, November 16-21, 1958; pp.917-935]

The Analogy between Mechanical Translation and Library Retrieval

M. MASTERMAN, R. M. NEEDHAM, and K. SPÄRCK JONES

ABSTRACT. Any analogy made between library retrieval and mechanical translation is usually made by assimilating library retrieval to mechanical translation. We desire to draw the converse analogy; that is, to assimilate mechanical translation to library retrieval. To do this, mechanical translation procedures must be generalised and made interlingual, until they become as general as library retrieval procedures already are. This generalisation can be made if the mechanical translation procedure is based on a thesaurus. The nature of a thesaurus is discussed in Section 3. This type of procedure has already been used for library retrieval, but not for M.T.; the use of a thesaurus for both fields enables a new, very general field to be exactly defined, namely the field of semantic transformation. This field would have application to library retrieval, mechanical translation, and probably also to mechanical abstracting. The purpose of this paper is to develop the application of this generalized procedure to mechanical translation, referring also to its use for library retrieval. For this purpose, an analytic examination of the translation procedure is required, as linguists object to the analogy that we are making by asserting that a library retrieval type of procedure will not translate syntax.

It is asserted that a generalised mechanical translation procedure cannot translate grammar and syntax as these do not correspond between different languages. There is a general answer, and a particular one, to this criticism. The general answer is that present procedures for translating between different pairs of languages generate such complexity that they do not form an adequate basis for future M.T. research. The experimental work done by workers in the U.S.S.R. is examined. The particular answer is that since recent mechanical translation experiments using a thesaurus show, contrary to expectation, that this method can interlingually translate semantic meaning, it seems not impossible that, again contrary to expectation, it can be used to translate syntax.

Such an extension is suggested by the linguist M. A. K. Halliday. He defines the syntactic operators of a source language in terms of a set of interlingual *questions.* This procedure is criticised.

M. MASTERMAN, R. M. NEEDHAM, and K. SPÄRCK JONES Cambridge Language Research Unit, Cambridge, England.

A generalised translation procedure, using a thesaurus, is related to the semantic problems of mechanical translation. A thesaurus is defined. Recent work done by the Cambridge Language Research Unit is described to illustrate this procedure. Experiments, done also in the C.L.R.U., using the same procedure for library retrieval, are described. The result, a conception of a procedure of generalised semantic transformation, is considered.

This semantic transformation procedure is extended to cover syntax. The questions used by Halliday can be turned into thesaurus heads. Some examples of interlingual translation of syntactic form are given. Research on these lines is continuing. If this method of generalised mechanical translation proves feasible, M.T. becomes straightforwardly an extended case of generalised retrieval.

Proposal to create a single general theoretic field of semantic transformation, with application to library retrieval and to M.T.

Many documentalists have insisted that there is an analogy between mechanisable procedures for retrieving documents and procedures used in mechanical translation (M.T.). The analogy between the two has usually been drawn, however, by assimilating library retrieval to translation; not the other way round. A coded library classification has been envisaged as an exact and interlingual library language. Any request for information, made in a particular language, must be translated into the interlingua, and also coded, if the retrieval procedure is to be mechanical (1).

We wish to draw the analogy conversely: that is, by assimilating interlingual mechanical translation to retrieval. Now, in the present state of research this analogy can only be drawn at all precisely between one form of library retrieval procedure, and one form of mechanical translation procedure; these two analogous procedures are those, in each field, which make use of a thesaurus. The proposal that an improved type of library retrieval procedure could be devised, using a thesaurus, of the type of Roget's famous *Thesaurus*, instead of a term classification, has already been made by American workers in this field (2,3). The proposal that semantic meaning can be translated using a thesaurus was first made by the Cambridge Language Research Unit (England), at the Second International Conference for Machine Translation (4,5,6). We propose, then, that a conceptually based, thesaurus type of language classification should be used for a completely generalised retrieval procedure, this classification procedure being, by its nature, interlingual. The development of this procedure makes possible the definition of a general theoretic field of semantic transformation. Of this field, a well-defined mathematical model can be made (7).

Surprisingly enough, the proposal that such a general field should be created seems far more revolutionary to mechanical translation specialists than to documentalists specialising in library retrieval. Translation specialists, and, in particular, linguists deny even the possibility of the analogy by maintaining that any classification of language based on a thesaurus can, at best, only hope to translate semantic meaning, whereas language is primarily a system of grammar and syntax; and both of these are notoriously monolingual. It could be said, indeed, that a library classification is like a non-grammatical language and that a thesauric library retrieval procedure could therefore hope to retrieve from it. But it is obvious, so the argument runs, that any mechanical translation procedure, before it starts dealing with subtle questions of semantic ambiguity, must deal with crude questions of how to translate grammatical and syntactic form; and these are both notoriously monolingual. Since, therefore, grammar and syntax cannot be translated by an interlingual thesaurus procedure, the analogy we wish to draw falls to the ground: it has no application to any procedure for mechanical translation.

The object of this paper is to refute this criticism by showing how a type of retrieval procedure, based on a thesaurus already being used for the experimental translation of semantic meaning, might also be extended so as to translate grammar and syntax. It is only by showing the procedure in action that we can hope to make clear what seems to us this most fundamental and important analogy between library retrieval and mechanical translation; we hope to show the nature of the generalised procedure by considering how it can deal with the particular problems of one of the fields in question, namely M.T. And this is all the more necessary in that the field of mechanical translation, unlike that of library retrieval, has not hitherto been approached at all from this point of view.

1. Application of the method to M.T.

On July 10th, 1957, M. A. K. Halliday read a paper, to the Cambridge Language Research Unit, and later in a developed form, to the International Congress of Linguists, held in August, 1957, in Oslo, in which, speaking as a descriptive linguist, he described a method which might be used to carry out an interlingual analysis of the syntax of a language (8,9). This method was nicknamed the Twenty Questions Method of Analysis.

Before discussing the method, however, we must give a provisional reply to those M.T. workers who deny the existence of an analogy between the mechanical translation and retrieval fields. These may ask, "Why attempt an interlingual translation between languages when we know that the grammar

and syntax of different languages do not correspond?" They may also ask, "Since it is mechanical translation of technical material which is urgently required in order to make scientific information more generally available, why not have, as the U.S.S.R. mechanical translation workers have, a set of two-language programmes, to translate from, e.g., Italian into French, or from Chinese into Russian, using for any particular text the appropriate programme?"

The answer to these questions, still keeping for the moment within the M.T. field, is that those who use such an approach, constructing a separate programme, to be stored by the machine, for every pair of languages, fail to consider the complexity which the method itself generates. Only one group of workers has extensively tried this method out: the Mechanical Translation Research Group of the U.S.S.R. Academy of Sciences. The project is described in an informative recent paper by I. K. Belskaya (10). This paper explicitly sets forth the restrictions on translation necessary to limit the complexities generated by the method itself. These are (1) *severe limitation of the input text:* only mathematical texts were used, the translation being from Russian into English; and the U.S.S.R. group only at present envisages mechanical translation of scientific texts; (2) *limitation of vocabulary:* in order to limit the number of multiple meanings required for successful dictionary entries, a separate entry was used for each whole word—the attempt to economise on storage space by dividing words into "chunks," or sub-words (11) was abandoned; (3) *multiplication of dictionaries:* different dictionaries were required for all the different fields, even when translating between the same pair of languages.

These experiments show that a mechanical translation programme constructed on the Russian model does not straightforwardly translate between two languages. What such a translation programme does, when used with, e.g., a technical mathematical dictionary and a general dictionary containing the common words of the language, is successfully to translate English mathematical texts into Russian. This is a tremendous technical achievement. But it is inadequate as a directive for future research. The failures, cited by Belskaya, of attempts by cryptographers and logicians to find a common basis, statistical or mathematical, to language, might indeed cause us to abandon the goal of interlingual translation. But we cannot abandon the attempt to achieve intertextual translation. If we cannot feed into a computer and translate, from a single source language, e.g., a novel, a philosophical treatise, a mathematical system and a botanical paper, without using separate programmes and dictionaries, we are not translating between pairs of languages. We are merely translating between pairs of texts. And mechanical translation on this basis is not a commercial prospect.

If we reconsider the Russian experiments, therefore, with the necessity for intertextual translation in mind, we are tempted to ask, "Can we at once have a more general approach to the problem?" This question seems all the more appropriate when we find that the U.S.S.R. group themselves think that a more general attempt to translate syntax might be successful. Belskaya says:

> Special experiments were made in order to find out whether the same grammatical programme can be applied to a text having as little to do with mathematics as, say, an article from *The Times*, or a page from Charles Dickens. These experiments proved the success of our ideas on the possibility of having a universal grammatical programme for the machine translation of any two languages. Our general principles have withstood another test: they were extended to cover machine translation from languages differing from English in structure as much as Japanese, Chinese, and German. These experiments having been successful, the principles (underlying the Russian grammar and syntax programme) may be considered as basic in the solution of machine translation problems.

Thus even the U.S.S.R. group, whose approach is strictly particularised and inductive, admit that there may be general ascertainable principles underlying the mechanical translation of grammar and syntax. The next object, then, of linguists associated with machine translation, ought to be the discovery and development of these principles, rather than further experiments on particular texts. We propose that this research should be pursued by substituting for the particularised methods of linguistic analysis at present in use among workers on M.T. the completely generalised methods at present in use in library retrieval; that these, having been given thesauric linguistic application, should be put on a machine, and the results examined. Such a method, which is essentially algorithmic and deductive, does not, of course, invalidate the step-by-step method of inductive generalisation, at present being used in U.S.S.R. But the light that it throws upon the whole process of semantic transformation, and the simplifications which can be attained by means of it, make it in our view a preferable basis for the next stage of research.

2. A suggested interlingual analysis of syntax

That M.T. research could be thus generalised is the opinion already of one linguist, M. A. K. Halliday. We must next, therefore, examine and criticise the method he suggests for the interlingual mechanical translation of grammar and syntax, before further considering the problem of whether fully interlingual and intertextual mechanical translation of scientific texts is possible. Halliday's method was first to make a strictly monolingual analysis of the

input language. He then made a further interlingual analysis of the language. For this interlingual analysis he does not recommend a generalised transfer grammar, of the kind developed by the American descriptive linguists, Z. Harris and N. Chomsky (12,13). He recommends using a more direct analytic method. This owes much to 19th century historical linguists. But Halliday's analysis, unlike theirs, is not evolutionary. First, he makes a rigid distinction between types of chunk, the operators of a language, and the arguments. (Roughly, the functions of operators are dealt with by grammar books; those of arguments, by dictionaries.) The operators are identified by their relation, positive or negative, to a number of categories (provisionally about 60). The arguments are then classified by referring to groupings of these systems (14).

Basically, therefore, Halliday makes first a monolingual grammar, and then an interlingual analysis of each language, the latter being quite distinct from the former. The monolingual grammar resembles those of descriptive linguists, except that it refers only to operators; the arguments are later defined by referring to the operators. The interlingual analysis, the key to the whole method, demands reference to extralinguistic contexts; only after these have been ascertained are the operators related to the arguments. The relation of any operator to the extralinguistic context is determined by asking questions, the answer to which can be "Yes," "No," "Both," "Neither." This procedure resembles that of the game "Twenty Questions," from which the method derives its name. The two methods differ, however, in that, for the linguistic analysis, in most cases, the answer to one question does not influence the next.

The interlingual analysis may proceed as follows. Take, for example, the French operator *la.* A normal grammatical description would classify this as either the feminine definite article, or the feminine accusative pronoun. We assume that *la* has already been subjected to a monolingual French analysis giving, e.g., gender. We now carry out the interlingual analysis: we do not ask "Does *la* belong to any gender system?" because it is notorious that the gender systems of different languages do not correspond. Therefore we simply ask: "Can *la* tell us anything about sex?" By this change of question we refer, not to the intralinguistic context (i.e., that of French), but to the far more general extralinguistic context (i.e., that of the human race divided into sexes). English has no genders, French has two, German three, Icelandic six, but English, French, Germans, and Icelanders alike fall into communities of only two sexes. Therefore the answer to our last question is "Yes." We may then ask: "Does *la* refer to animate or inanimate objects?" The answer is "Both." To the question "Does it apply to present or non-present time?" the answer is "Neither." And so on.

Now it is clear that, even from the pure linguist's point of view, Halliday's suggestion is of great research interest, since what he proposes is to use the precise and elegant analytic methods of contemporary linguistics to analyse, both monolingually and interlingually, the context grammar of particular texts. (These analytic methods, as is known, depend on being able to break up the older grammatical units, such as noun, verb and the rest, into weaker but more precisely definable units, special to each language, from which, by referring to the intralinguistic context-grammar of a text, the older type of unit, can, where it is required for that particular language's analysis, be built up.) In order to extend this method to make it apply to an interlingual grammar based on extralinguistic context analysis, it is evident that Halliday must take seriously the analogy, to which older linguists have paid nothing more than lip service, between intralinguistic context and extralinguistic context, and the way that each might be used to build up grammar and syntax. And, from the pure linguistic point of view, this is a very interesting thing to do. But if we consider his interlingual analysis from the point of view of mechanical translation rather than from that of linguistics, it is clear that it has serious defects. These are (1) that the monolingual analysis is too complicated a way of obtaining the list of operators of an input language; a first approximation to these could be obtained with far less trouble by consulting a grammar book, and then, by applying the procedures, to find out where and why the translation had turned out wrong; (2) that though the method *analyses,* it does not *translate.* For mechanical translation purposes it must be turned from a method of analysis into a translating procedure; (3) that the method is essentially not linguistic at all, but logical. Therefore logical sophistication, rather than linguistic scholarship, should be used to make the question system more economical.

3. *A procedure for the translation of semantic meaning, using a thesaurus*

This bringing to bear of logical methods on problems of M.T. is at present being tackled by only one unit. Only the Cambridge Language Research Unit uses logical methods together with linguistics for mechanical translation research. It is no coincidence, therefore, that it is the only unit which is simultaneously investigating procedures for mechanical translation, library retrieval, and mechanical abstracting. Although it is almost universally assumed, by mechanical translation research groups, that it is the linguist, not the mathematician, who provides the computer programmer with the data for the translation programme, we contest this. We consider that the very nature of the

problem of interlingual mechanical translation is like that of information retrieval in that it demands a general, that is, a logical approach. Belskaya considers that as no logical system for interlingual translation has yet been devised, none could exist. We hold not only that such a system can exist, but also that it does exist as soon as the output language is analysed, not as a dictionary but as a thesaurus. The interlingual system required for mechanical translation and library retrieval alike is thus not a new interlingual language. It consists of a logical system giving the structural principle on which all languages are based. This principle is that language, seen interlingually, consists of an ordered finite set of clusters of synonyms (the synonyms being, of course, different for each language), which can be represented by a corresponding ordered set of topics, or very general abstract nouns, or heads. These heads are homogeneous, that is, they do not themselves divide up into different parts of speech, since the synonyms of which they represent the common principle of synonymy will be in different parts of speech in different languages. They are vague; their "meanings" cannot be given except by reference to the sets of synonyms in any language which represent them, and these sets of synonyms are not precisely bounded. They are unobservable; some existing word in any given language may be usable, in an extended sense, to represent some idea common to a set of synonyms, or it may not, in which case, either a new word has to be invented, or the set has to be left identifiable only by reference to its position in the total ordered set of heads. Thus, in English, the head-words "greatness," "smallness," "region," "base," "land," do exist; the head-word "materiality" does not. In short, these new interlingual units of semantic transformation have a series of theoretic "ineffable qualities" attached to them just as Newton's infinitesimal operators seemed to have to his contemporaries in 17th century philosophy and science. But like Newton's operators, these units also can be used in determinate mathematical procedures; such a procedure is given in detail, for mechanical translation, in Appendix 1, and for library retrieval, in Appendix 2.

Thus the nature of a thesaurus, or general ordering principle for language, may be briefly characterised as follows. A thesaurus, or synonym dictionary, e.g., *Roget's Thesaurus,* unlike an ordinary dictionary, consists of an ordered set of lists of synonyms grouped under a comparatively small number of concepts, or topics, or heads. (We use the word "head" to describe these because it is the word which Roget himself used.) These heads are themselves arranged in a single or multiple hierarchy, usually in decreasing order of generality; thus the chapter of contents of a thesaurus, taken by itself, will exemplify the mathematical system of a tree. The whole thesaurus cannot be taken as a tree, however; because, in it, the words of a language will always occur more than

once; a synonym occurring under any given thesaurus head represents only the use of that word in that language in that context. Thus the occurrence of the English word "plant" in *Roget's Thesaurus* under "Agriculture" signifies that "plant," used agriculturally, means "something growing." The occurrence of the same word "plant," under "tool," signifies that "plant" here means, "total engineering apparatus"; and so on. The head-words of the thesaurus do not reoccur; but the synonym words given under them do; and this means that the total thesaurus, consisting of chapter of contents, numbered list of heads, and lists of synonyms, cannot, mathematically speaking, be represented as a tree, but must be represented as a lattice; that is, as a partially ordered set of which any two elements will not only have a point in common above them, higher in the hierarchy; but also a point in common below them, lower in the hierarchy. The advantages for library retrieval of substituting a lattice for a tree are exemplified in Appendix 2; the advantages for mechanical translation are exemplified in Appendix 1 by the fact that the mechanical translation procedure does obtain an output. Thus the theoretic importance of making the new field of semantic transformation work on a lattice, rather than on a tree is that the lattice, unlike the tree, guarantees that translation-points, and retrieval points, in the system do exist; that is, that, by using the system, information can be retrieved, and translations obtained.

In practise, both for library retrieval and for translation, it is so important to be able to locate the end points of the semantic transformation procedure that the thesaurus itself is always used in conjunction with a cross-reference dictionary. In *Roget's Thesaurus* the thesaurus itself occupies only the first half of the book; all the second half is occupied by a cross-reference dictionary in which those words of the English language which occur in the thesaurus are listed alphabetically, each word being followed by a list of the numbers of the thesaurus heads in which the word occurs. Now in *Roget's Thesaurus* itself the cross-reference dictionary is of course unilingual; *Roget's Thesaurus* is normally used only by authors for improving their style, that is, for translating from English into English. When interlingual translation is contemplated, however, the cross-reference dictionary, for any source language, must be bilingual; that is, its use must transform the chunks of any input text into sequences of lists of thesaurus heads (see Appendix 1). Thus, although interlingual translation is contemplated, for each source language there must be a separate cross-reference dictionary. The lattice of thesaurus heads will be interlingual; but the lists of synonyms, idioms, and paraphrases appearing under any given head will be individual to any output language. Thus each output language must have its own lists of synonyms to fit under the heads of the interlingual thesaurus. And the process of transforming the sequence of

source-units (terms in the request, in the library case, chunks in the source text, in the translation case) into sequences of sets of thesaurus heads; of operating an algorithm to select from among these heads and/or to substitute for them other heads or sets of heads from the total thesaurus; and of transforming the selected sets of heads into output (documents or sub-documents, in the library case, synonyms in common between the selected heads, in the translation case), this total process constitutes the process of semantic transformation, and the total possible field to which it can apply the proposed new general theoretic field.

It is worth remarking that, when this procedure is used, translation, like retrieval, becomes irreversible and asymmetrical. The words of the source language must be divided into as fine sub-words as possible, so that the dictionary entry of each chunk shall give a whole list of head-numbers; this list, if it is full enough, exactly defines the spread of that chunk's ambiguity, and distinguishes this spread from the ambiguity-spreads of cognate chunks. Analogously, in the library case, the library-user's request, if it is at all complex, must be analysed as finely as possible into terms. On the other hand, the synonym lists, in the output language, must be as long, complete, and vivid as possible, consisting not only of whole words, but also of whole phrases, sometimes even of whole sentences. Analogously, in the library case, the output consists of a series of whole documents; or, at the least, of references to sections or paragraphs within them.

To recapitulate: a thesaurus, as Roget himself saw (15), is not primarily an analytic tool; it gives a procedure for finding analogies; that is, for finding relevant information; that is, for finding translations. It is organised on a hierarchical principle which is like that of a library-retrieval tree classification, except that, instead of forming the mathematical system of which the model is a tree, it forms the mathematical system whose model is a lattice. It is interlingual, in the sense that the heads have synonyms in any language. It gives a general solution for semantic problems; that is, those arising from the unusual use of words or from multiple meaning. It deals therefore, and by a single procedure, with the most difficult problems facing alike mechanical translation research, and research into methods of mechanising information retrieval, and methods of mechanising the process of sub-titling and abstracting. Of course, information retrieval is a perfect field for applying a thesaurus procedure, just because a library system may be regarded as a language without syntax. But we claim to have shown also that the use of a thesaurus has immense possibilities also for mechanical translation itself. In experiments performed at the C.L.R.U., a thesaurus procedure has been used *(a)* to translate a novel use of a word in an Italian scientific paper, *(b)* mechanically to translate a line of Latin

poetry, (c) to retrieve documents from a library, and (d) mechanically to construct the essentials of discursive paragraphs of text (7). In all of these, the thesaurus has been used only to translate semantic content. So we return once more to the question, "Can it also translate syntax?"

4. *The same procedure, related to syntax*

We already have the method devised by Halliday to analyse syntax. His type of questions, for instance, were used to analyse the Latin sentences, "Magnam multitudinem vidit" and "Ad ludum ambulamus." (The actual questions used were those obtained by M. Masterman and K. Sparck Jones.) We now have to ask, can a thesaurus procedure derived from these be used to translate? For all Halliday's questions can be rephrased as single words; these, in turn, can be replaced by thesaurus heads; and these, by their nature, will yield an English output, when the thesaurus-lists are in English. In principle, therefore, the problem of using Halliday-derived heads for translating instead of analysing can be solved, at a stroke, simply by turning his questions into heads. In practise, however, the general problem of interlingually translating syntax can be resolved into two difficulties: (1) Can the information given by the monolingual analysis of an input text (as done, e.g., by Mukhin (16), Richens (17), and Halliday be picked up by a dictionary leading straight to a thesaurus? Can the entries for such queer chunks as -AT- and -US of the Latin OB-STIN-AT-US lead into combinations of thesaurus heads? The answer is that in some cases they can, though we do not know if this is true of all cases. (2) Can a thesaurus really be used to translate grammar as well as syntax? That is, can grammatical form, for example, the subject-predicate relation, be treated as semantic information leading to a thesaurus? The answer to this again is, in some cases, yes. Whether such entries can be constructed for all features of monolingual grammar, we do not yet know. What is already evident is that answers to these two questions can only be obtained if the principles of monolingual analysis are reconsidered from the fundamentally different viewpoint of a thesaurus-maker; and this means approaching the problem from the viewpoint, no longer of linguistics, but of retrieval. And this is so because a thesaurus is essentially a logical structure, designed to retrieve relevant information from an antecedently constructed complex, namely a thesaurised language or sub-language.

REFERENCES

1. R. A. FAIRTHORNE, *Patterns of Retrieval.* Amer. Doc. 1956.
2. C. M. MOOERS, Information Retrieval on Structured Content. Paper of Third London Symposium on Information Theory, 1956.
3. H. P. LUHN, A Statistical Approach to Mechanised Literature Searching *I. B. M. Journal of Research,* 1956.
4. M. MASTERMAN, *The Potentialities of a Mechanical Thesaurus,* C.L.R.U.
5. A. F. PARKER-RHODES, *An Algebraic Thesaurus,* C.L.R.U.
6. M. A. K. HALLIDAY, Linguistic Basis of the Thesaurus-Type Mechanical Dictionary and Application to English-Preposition Classification. Papers read at Second International Conference on Mechanical Translation, M.I.T. October, 1956. *Abstracts in Mechanical Translation 3,* [2]. See also, for Halliday's paper, *Mechanical Translation, 3,* [3].
7. A. F. PARKER-RHODES and STEPHEN WHELAN, Appendix to *Information Retrieval and the Thesaurus,* by R. M. Needham and T. Joyce; paper presented to Programme Committee, Area 5, *International Conference on Scientific Information.*
8. M. A. K. HALLIDAY, The Linguistics of Mechanical Translation, *Proceedings of the 8th International Congress of Linguists,* Oslo, August, 1957.
9. M. MASTERMAN, Linguistic Problems of Mechanical Translation, to be published in forthcoming issue *of Mechanical Translation.*
10. I. K. BELSKAYA, Machine Translation of Languages, *Research,* Oct. 1957.
11. R. H. RICHENS and M. A. K. HALLIDAY. Word Decomposition for Mechanical Translation; 8th Meeting of Linguists, Georgetown University 1957.
12. Z. N. HARRIS, Transfer Grammar, *International Journal of American Linguists, 20,* [4], 1954.
13. N. CHOMSKY, *Syntactic Structures Ianua Linguarum,* 4'S Gravenhage, 1957.
14. M. A. K. HALLIDAY'S working charts (unpublished).
15. *Roget's Thesaurus,* author's preface, 1931 edition, Longman's.
16. I. S. MUKHIN, *An Experiment on the Mechanical Translation of Languages carried out on the B.E.S.M.,* published by the U.S.S.R. Academy of Sciences, Moscow, 1956.
17. R. H. RICHENS, General Programme for Mechanical Translation between Any Two Languages via an Algebraic Interlingua Paper read at 2nd International Conference on Mechanical Translation M.I.T. Oct. 1956. (Abstract in *Mechanical Translation, 3,* [2].
18. R. M. NEEDHAM and T. JOYCE, The Thesaurus Approach to Information Retrieval, *Am. Document.,* July 1958.

APPENDIX 1

The contents of this appendix are taken from Margaret Masterman's paper "The Potentialities of a Mechanical Thesaurus."

The parts of the paper referred to are concerned with (1) the translation procedure, and the example, the translation of an Italian paragraph, used to show the procedure; (2) the discussion of particular difficulties which arose due to the unusual use of words.

The translation procedure is concerned, not with free *words*, but with *chunks.* A chunk is defined as "the smallest significant language-unit which can exist in more than one context, and which, for practical purposes, it pays to insert as an entry by itself in an M.T. dictionary" e.g., the Italian free word PIANTATORE is broken up into PIANT - AT - ORE. Each chunk, forming an entry in the M.T. dictionary, can have a number of meanings, or uses.

The range of uses of any chunk in a language, can be so envisaged that it forms a tree, the total dictionary entry of the chunk forming the point of origin of the tree. When any such a tree is connected, for translation purposes, with the corresponding tree in another language, the two trees together form a lattice each point of which looks both ways and is itself a translation point. (For an amplified discussion of trees and lattices vide Margaret Masterman "Fans and Heads," and "Outline of a Theory of Language," Work-papers, C.L.R.U.)

A point on a lattice, or a multilingual dictionary article, is analogous to a topic, or *head* in a single language *thesaurus.* "Discussion of this analogy led to the suggestion that a multilingual M.T. programme might be developed (given an imaginary computer of indefinitely expandable size) in which the multilingual dictionary might be replaced by a target language thesaurus."

A brief account of the programme which was developed and the thesaurus using translation experiment which was carried out on an Italian paragraph, is given below. At this point the procedure was only applied to semantic heads; no attempt was made to analyse syntax by the use of a thesaurus.

THE PROGRAMME

In the programme three operations are carried out on the input text which has been broken up into chunks:

1. The chunks are matched with the chunks of a pidgin dictionary, giving a pidgin English output.

2. The pidgin chunks are matched with the relevant entries in the cross reference dictionary of a thesaurus. *(Roget's Thesaurus,* with additions, was used in this test.) This stage gives an output of thesaurus heads relevant to a greater or lesser degree to the final translation output required.

3. A number of operations, restricted by rules, select from the list of thesaurus heads given under the entries in the cross reference dictionary the one which is appropriate. (This is not the final selection, as under each thesaurus head there is a list of synonyms; this problem was not, however, dealt with in this paper.)

A. *Chunking of the Italian passage*

Each chunk was written on a card which was used for the matching process:

LA PRODUZ-ION-E DI VARIET-A DI PIANT-E PRIV-E DI GEMM-EASCELL-ARI O PER-LE-MENO CON GERMOGL-IA SVILUPP-O RIDOTT-O, INTERESS-A DA TEMPO GENET-IST-I ED AGRONOM-I, TAL-E PROBLEM-A SI PRESENT-A PARTICOLAR-MENT-E INTERESS-ANT-E PER ALCUN-E ESSENZ-E FOREST-AL-I E FRUTT-IFER-I, PER LE PIANT-E DI FIBR-A, MA SOPRATUTTO PER IL TABACC-O, IN QUEST-A COLTUR-A E INFATTI IMPOSS-IB-IL-E MECANIZZ-ARE L'-ASPORT-AZION-E DEI GERMOGL-I, ASCELL-ARI, NECESS-ARI-O D'-ALTRA-PARTE PER OTTEN-ERE FOGLI-E DI MIGLIO-E QUALIT-A.

B. Matching of these chunks with English pidgin chunks using the Italian-English pidgin dictionary

The type of entry in the dictionary is as follows:

AL-	...	-Y
FIBR-	.. .	FIBRE
I	...	THOSE-WHICH-ARE
GENET-	. . .	GENETIC

This matching gave a very pidgin translation. This was improved by using a syntax lattice procedure which gave synthesis routines, and the following pidgin translation was obtained:

THE PRODUCE-MENT OF VARIETY-S OF PLANT-S WITHOUT AXIL-ARY BUD-S, OR AT LEAST WITH SPROUT-S AT REDUCED DEVELOPMENT-S, INTEREST FOR SOME TIME PAST GENETIC-IST-S AND AGRICULTURE-IST-S. SUCH PROBLEM-S SELF-PRESENT PARTICULAR-LY INTEREST-ING FOR SOME FOREST-Y AND FRUIT BEARING ESSENCE-S, FOR THE PLANT-S OF FIBRE-S, BUT ABOVE ALL FOR TOBACCO. IN THIS CULTIVATE-URE IT BE IN FACT IMPOSSIBLE TO MECHANISE REMOVE-MENT OF ALL THE AXIL-ARY SPROUT-S, ON THE OTHER HAND NECESSARY FOR TO OBTAIN LEAF-S OF BETTER QUALITY-S.

This translation obviously fails at some particular points:

ESSENCE-S for ESSENZ-E,
SPROUT-S for GERMOGL-I,
SELF-PRESENT for SI PRESENTA,

strictly also ASCELL- should have been translated by the vernacular ARMPIT-.

These cases were examined in detail, by using the next stages of the procedure: the thesaurus cross-reference dictionary and the thesaurus. The three cases examined are important in that they represent the unusual uses of words:

ESSENZ-E is being used in a new way.
GERMOGL-I is being used technically.
SI PRESENTA is being used idiomatically.

Therefore the pidgin output must be retranslated.

Roget's Thesaurus was used in the normal way, i.e., the chunk of word in question is looked up in the cross-reference dictionary, e.g., *bud* gives:

(*a*) bud, head no. 367	*(e)* expand 194
(*b*) beginning 66,* 129*	(*f*) graft
(c) germ 153	*(g)* -from 154
(*d*) ornament 847*	*(h)* -dy 711 890

Cross-references which are asterisked are additions made to Roget so that it is made multilingual, i.e., to ensure that there is a reference corresponding to each chunk of the input text when turned into pidgin. These additions were legitimately made by comparing the synonyms given under related heads to discover those in common. It was necessary as in some cases the list of cross-references given in the dictionary was inadequate, and the thesaurus can only be properly used if a chain of meanings can be followed through the thesaurus. (It is sometimes possible to obtain the required meaning by following up the synonyms given under the heads without making any addition to the list of cross-references.)

C. *Particular cases examined using the thesaurus technique which represents the next stage of the procedure.*

(i) ESSENZ-E . . . ESSENCE-S

If the chunks FOREST AND FRUIT-BEARING ESSENCE-S are matched with the chunks in the cross-reference dictionary of the thesaurus, the following output is obtained:

forest Head no.	*57, 367, 890
and Head no.	37, 38
fruit Head no.	result 164
	produce 161
	food 298
	profit 775
	forbidden- 615
	reap the -s 973
	-tree 367
	fruitful 168
	fruition 101
	fruitless 169, 645, 732
bearing Head no.	relation 9
	support 215
	direction 278
	meaning 516
	demeanour 692
	-rein 752
	*fruit- 168, 637, 367
	*child- 161
essence Head no.	5, 398
essential	intrinsic 5
	*meaning 516
	great 31
	required 630
	important 642
	essentially 3, 5
	essential stuff 5

Operation 1

Pick out all the numbers which occur more than once in the above output; these are called ring numbers. The following is obtained:

Ring number	*Thesaurus head*	*Source of ring number*
367	Vegetable	Forest, fruit
161	Production	Fruit, bearing
168	Productiveness	Fruit, bearing
516	Meaning	Bearing, essence
5	Intrinsicality	Essence

Operation 2

Pick out all the ring numbers and order in a scale of descending frequency of occurrence. This gives:

5, 367, 161, 168, 516.

Operation 3

Compare, in twos, for common elements, the ring number thesaurus heads which represent the meanings of the pidgin chunks. A comparison, or intersection operation on two thesaurus heads is permitted only if there is some relation between the chunks from which the heads are derived. In this test the relationship is determined by the syntax lattice: an intersection is permitted only if the points on the syntax lattice corresponding to the chunks have an inclusion relation between them. Thus it is permitted to intersect two heads from the same chunk as this is a trivial inclusion relation. We also intersect the heads from different chunks provided there is an inclusion relation between the points representing the chunks on the syntax lattice. The direction of the inclusion relation determines for which of the chunks the output of the head intersection is to be taken as a new translation. The rule is that the output retranslates the lower of the two chunks, i.e., the one included.

In the case of any two chunks, *A* and *B*, the operation of comparison is called $A \cap$ B, and we have an ordering as follows:

$$A \cap A = A \equiv A \geq A$$
$$A \text{ covers B}$$
$$A \cap B = -B \equiv A \geq \text{B}$$

Carrying out this operation we get an output of which selected examples are given below:

$A \cap A = A \equiv A \geq A$	*Chunk comparison*	*Ring number comparison (in all possible pain)*
	Fruit ∩ fruit	367 ∩ 161
	Fruit ∩ fruit	161∩ 168
	Fruit ∩ fruit	367 ∩ 168
A covers B	Fruit ∩ bearing	161 ∩ 168
	Fruit ∩ bearing	161∩ 367
	Fruit ∩ bearing	161 ∩ 516
	Fruit ∩ bearing	168 ∩ 367
	Fruit ∩ bearing	168 ∩ 516
	Fruit ∩ bearing	367 ∩ 516
A ∩ B = B≡ ≥ B	Forest ∩ essence	367 ∩ 5
	Forest ∩ essence	367 ∩ 516

Further operations can be carried out by combining the chunks. The comparison Forest ∩ fruit cannot be carried out as their lattice positions are not inclusive; in this case, by chance, no intersections are prohibited, as all possible combinations of the numbers are made by other chunks.

Operation 4

List the common elements (or words) given by the intersection of the thesaurus heads; i.e., those words common to the lists of synonyms given for the separate heads.

e.g.,

Ring numbers	*Thesaurus heads*	*Output: Words common to the lists of synonyms for the heads*
5 ∩ 161	Intrinsicality ∩ production	Flower, etc., Head no. 22 (prototype)

This gives a new thesaurus head; it is necessary to carry out the intersection procedure, as under Operation 3, using this new head, with results as follows, e.g.,

5 ∩ 22	Intrinsicality ∩ prototype	example, specimen
516 ∩ 22	Prototype ∩ meaning	prototype, example

Operation 5

We wish to obtain alternative translation for a particular chunk, i.e., to select from a list of synonyms for the chunk a more appropriate word. The synonyms are obtained by applying the output words given by Operation 4 to the intersections of Operation 3. So that we get, e.g.,

Fruit ∩ fruit	367 ∩ 161 (Production ∩ vegetable)	flower

Therefore the final output, flower, is a retranslation of fruit. Similarly for ESSENCE-, the example word selected, we get:

Essence ∩ essence	5 ∩ 516 (Intrinsicality ∩ meaning) and then operations using the new head, 22 (prototype)	example

Example is therefore a retranslation of essence.

In the case of operations on two different chunks the synonym refers to the chunk which comes lower in the lattice.

A number of restrictive rules are required to regulate the final output.

(i) referring to Operation 3, $A \cap A = A \equiv A \geq A$.

When $A \geq A$ is taken, and yields $A \cap A$, if A is not a chunk but a ring number, take the output which is identical with the original chunk.

(ii) If the above rule operates reject all other output.

(iii) When selecting the final output, take the longest output first, i.e., if there is a synonym output for Fruit-bearing essences, prefer it to a synonym for Fruit-bearing.

Using these rules we get the final synonyms, as follows:

for FOREST ESSENCE we get forest flower

for FRUIT-BEARING ESSENCE we get fruit-bearing example,

(ii) GERMOGL-I . . . SPROUT-S

(iii) SI PRESENTA . . . SELF PRESENTS

Retranslations for these output words were carried through in exactly the same way as for the first example. The third test failed, however, as to retranslate it adequately the syntax of the whole sentence would have to be taken into account, and the syntax lattice was not developed to this point. Nevertheless, the truncated procedure yielded the interesting translation, “strike one as.”

CONCLUSION

A number of conclusions were drawn from this test which indicated that further work on thesauri for translation purposes might be fruitful.

Claims are made for the thesaurus procedure as following:

(a) It is a procedure for giving an idiomatic translation.

(b) It is possible to see where it goes wrong.

(c) The test gives useful information on the construction of a thesaurus; this would assist the making of a thesaurus directed to M.T.

(d) The only dictionaries used are the bilingual pidgin dictionaries. The major lexical emphasis is on the target language thesaurus; and this one thesaurus serves for translation from *all* languages into the target language.

(e) The procedure uses previous M.T. results, which show the efficiency of mechanical pidgin; at the same time further analysis of the input language is possible.

Difficulties of a computer holding a thesaurus might be solved by encoding the thesaurus in the form of lattices, the points of which represent chunks.

A number of modifications have since been made in the procedure developed in this test, although the idea of using a thesaurus has been continued.

1. The preliminary translation into pidgin was abandoned.
2. The syntax lattice, in the form used in the test, was also abandoned.
3. The matching process was revised.

The successive intersection process has been seen to be uneconomical. The point has been discussed in the Unit's papers on Information Retrieval, and in detail in "A Note on a Property of Finite Lattices" by R. M. Needham.

When the original paper was written the problems of analysing syntax were considerable and were thought unconnected with the thesaurus procedure for semantics.

APPENDIX *2*

The brief account given below is taken from those parts of "The Thesaurus Approach to Information Retrieval," by R. M. Needham and T. Joyce, which refer particularly to the use of a thesaurus.

The problem of library retrieval is to describe documents so that, for any request in ordinary language, all relevant documents can be retrieved by a simple operation, without losses or irrelevancies.

If a large number of terms are used to describe a document, the existence of synonyms is likely; in a system such as Uniterm no attempt is made to bracket the synonyms, which means that a request will produce only the document described in identical terms and not in synonymous ones. If the existence of synonyms is avoided, by using a small number of exclusive descriptors, the description of a document in terms useful for retrieval is more difficult, also it is equally difficult to relate a request to the descriptions of documents. A further difficulty is that descriptions only list the main terms, and take no account of their relations to one another. The C.L.R.U. experiments being carried out make use of a thesaurus, a procedure through which it is hoped that these difficulties will be avoided, and that a request for a document

although not using the same terms as those in the document, will produce that document and others dealing with the same problem, but described in different, though synonymous, terms.

PROCEDURE

1. Term abstracts are made of the documents; the descriptors of a particular document are thus terms taken from it. It is at this point that the problem of synonyms must be solved without making the description procedure too rigid. The solution is arrived at in the next stage of the procedure.

2. The terms are then arranged so that near-synonyms are accommodated. This can be done by introducing a partial ordering relation, in which more specific terms are included in more general ones dealing with the same topic. So that in making a request for *A,* we are given B if the relation $A \geq B$ holds This makes allowance for loose description and also for structure.

This ordering is in effect making a thesaurus; each term can from this point of view be described as a head, and the inclusion relations of the terms correspond to the general-specific relations of a group of thesaurus heads, if we discuss the thesaurus in the terms of the ideal rather than the actual. Similarly, the parallel does not quite hold if we consider the list of synonyms given under each head in a thesaurus such as Roget, although it could be made to do so. The synonyms given under each Roget head are not ordered in any kind of relationship except that of being synonyms, or rather, near synonyms, for the head. If we consider the retrieval terms, however, this situation does not exist: the ordering relations are much more fully worked out, so that what would be synonyms in the thesaurus appear as terms either including, or more usually, included in, the original term; only synonyms in the strict sense are equal. Usually the near-synonyms appear as subordinate to, or more specific terms than, the main term. In this way the existence of near-synonyms is allowed, so that there is no loss of information; at the same time loosely expressed requests can be made precise, it being possible to treat each term if necessary as independent. The hierarchy constructed also made it possible, in dealing with requests in the retrieval procedure, to obtain a scale of relevance in order to secure the correct output.

The partially ordered set is converted, by including latent elements, into a lattice, making the ordering of the terms more systematic, and also making the actual retrieval procedure carried out on punched cards (for details see the paper) easier.