[From: *Information retrieval and machine translation*, ed. Allen Kent. International conference… September 6-12, 1959, Cleveland, Ohio; part 2 (New York: Interscience Publishers, 1961)]

CHAPTER 39

Some Recent Work on Thesauric and Interlingual Methods in Machine Translation

A. F. PARKER-RHODES

Cambridge Language Research Unit, Cambridge, England

I. THE TRANSLATION PROBLEM

The aim of research on machine translation is to devise a fully automatic method of effecting translation of a given text in a source language into another text in a target language. Success will have been achieved if the second text, the output, is acceptable, by the standards commonly applied to man-made translations, as "a translation" of the input. It is therefore first necessary, in a general discussion of the problem, to decide what we mean by "a translation." This is more difficult than it looks at first sight. It is not for example adequate to postulate that the output should convey the same information as the input (unless we widen the meaning of "information" a good deal beyond its usage in information theory); for we shall not accept as an accurate translation a rendering which gives the same factual information with quite different emphasis. Thus, "we reached the town no more than an hour after dark" and "we did not reach the town till an hour after dark" are by no means equivalent to each other; they will not usually be equally acceptable translations of any original. But it would require a very sophisticated idea of "information" to make them appear different in information content.

The difference between these two sentences resides in the difference between the forms "no more than" and "not… till"; it is therefore necessary (and probably sufficient) in rendering either into a second language, to find in the latter a sufficiently close rendering of each of these forms. They need not, however, take the form of different words or phrases; in some languages the distinction can be made by syntax or word order alone. The fact that parts of the meaning of a sentence can be conveyed either by choice of words (i.e. lexically) or by manner of combination of words (i.e. syntactically) according to the language one is using is just one of the complications which beset the attempt to reduce the relationship of "translation-equivalence" to mathematical terms. And to be able to do this is a

necessary preliminary to any automatic translation procedure which is to be applicable with any generality (i.e., to any type of text in any pair of languages).

The meaning of any text is in fact carried partly by the lexical content of the words employed, and partly by the structure in which they are put together; but not all the lexical content conveys meaning in any given text, nor does all of the structure require to be reproduced on translation. Thus, the meaning of "the table on page 13" does not in any way depend on the fact that (in English) "table" can also mean a piece of furniture, "on" can mean "concerning", and "page" can mean "serving-boy"; nor does the "noun-preposition-noun" structure have any validity in other languages, even though it does affect the meaning in English (compare "the page on table 13" !).In Chinese, for instance, the appropriate structure would be of the type "no. 13-page's table". We are therefore faced at the outset with the problem of defining how much of the structure and how much of the lexis matters i.e. carries meaning required for translation; this has been called by Mooers, in connexion with information retrieval, the problem of "structured content." It is basic to translation theory.

II. STRUCTURED CONTENT

The practical problem is to ascertain, for a given input text, what is the minimum relevant structure and the minimum relevant semantic content at each point in this structure, and to embody this essential minimum in a suitable form for subsequent processing, i.e. to encode it intelligently. The encoding problem is easily solvable where there is no structure to contend with, and gets progressively more difficult, the more complex the structure; it follows that our aim must first be to transfer as much as possible of the information which the input text carries in its syntax over to the semantic side. For instance, in English when two nouns are juxtaposed, as in "horse race" or "bedroom furniture," the first is being used as a qualifier of the second; this is a necessary piece of information carried in a syntactic construction, but it could be carried in an explicit unit, i.e. extra word, inserted into the text. This extra word could be regarded as an adjective-forming suffix attached to the first noun. The result would be a three-unit structure instead of a two-unit one, in which bracketing-pattern is important, but not word-order.

By devices of this type we can therefore reduce the amount of structure needed. The result of our efforts will be a coded form of the input text. The code used is called an "interlingua," and in our over-all translation procedure it will appear as a middle stage, to which we first reduce the input, and out of which we thereafter construct the output. It will therefore need to be subjected to a variety of computational operations. These operational requirements of the interlingua constitute in fact an important limitation on the extent to which we can eliminate structure from it.

The most important point is that the units, of which the structure

is built up (i.e. the "words" of the interlingua), must be susceptible of mathematical operations, and must therefore be interpretable as elements of some tractable algebra. The operations we shall have to perform have the following objects: (i) to adjust the meaning or range of meanings represented by each unit to accord with the requirements of the context, and (ii) to identify in the output dictionary whatever word of the target language most nearly represents the "meaning" encoded by the interlingua for any given unit, and to assess also how nearly this meaning is conveyed. There are no doubt many types of algebra in which representations of these operations can be set up, but in our work we have been impressed by the especial suitability of lattice algebras for this purpose. We find that we can in fact represent most of what we have to do in terms of the operations of lattice algebra, and that if we can use certain types of lattice (the "canonical lattices" of CLRU publications) these operations can be performed with very great speed by electronic computers; some additional aid from outside lattice theory will still probably be required however; it has been suggested by M. Masterman that combinatory logic might successfully fill the gap.

If the units of our interlingua are to be representable as elements of a lattice, they must be themselves without any internal structure. This requirement puts out of consideration certain types of interlingua which we have been previously working on (for example, the system called "Nude" (CLRU/ML76)). Another requirement is obviously that we want to give the code the greatest possible variety for a given amplitude of (i.e. number of bits in) the symbols used. This also affects the construction of the interlingua quite materially. Let us now consider how these ideas work out.

III. INTERLINGUAL FORMULAE

The principle on which interlinguas are constructed is essentially that their units consist of formulae, made up of terms and connectives. The terms convey the semantic content which the formula is to represent and the connectives provide whatever structure has to be imposed on this content. As we have seen, our formulae have to have no internal structure, which means that we can use only one connective, and this must be commutative and associative, so that bracketing and order of terms are irrelevant. This imposes straight away a limitation on the variety of the interlingua as a code; that is, the range of meanings which a single formula can convey is restricted.

Consider the ordinary meaning conveyed by the English word "perforate." This can be paraphrased as "insert small holes in." There is structure here of a sort: it is the holes that are inserted, and not the holes that do the inserting. If we used two sets of terms one for agents and another for objects, we could render the required idea in, say, the following form: "make.internal.small (obj). .internal (obj).gap (obj)" and the order and distribution of the component terms, all of which are of the degree of generality which experience shows to

be appropriate for our purpose, would not matter (because all the terms marked (obj) could be brought together by a fixed rule). But to apply this method consistently throughout the whole system would result in a great increase in the number of terms required.

If on the other hand we refuse to duplicate terms to this extent, we must recognise the structure inherent in the notion of inserting holes (verb-object relation in grammatical parlance) in some other way. The alternative is to regard "perforate" as a compound notion, to be translated in our interlingua by two formulae, between which a stronger structural relation can be postulated. Thus, the dictionary entry for this word, on this scheme, would consist of a sequence of two formulae, say "small.internal.gap" followed by "make.internal." In general we should have to postulate also some specific connective between such a pair, but as we shall see we can manage with only one interformular connective, so that it need not be mentioned in the dictionary entry.

This expedient however must not be too heavily indulged in either, or we shall find that every input sentence becomes unmanageably long when translated into its interlingual form. The ideal is to multiply formulae as little as possible, subject to the condition that we shall never have to translate one interlingual formula by a compound expression in the target language (except to the extent that it is reasonable to include a few such expressions, of an idiomatic character, in any dictionary). We are therefore faced with finding a compromise between minimizing the number of terms used to construct the formulae, and maximizing the variety of the code which the resulting formulae provide us with. Only wide empirical experience can lead to a final balance in such a matter, and needless to say we make no claims to have achieved this yet.

The formulae we use will then be constructed from terms drawn from a limited repertory, and subject, we may expect, to certain restrictions on what combinations constitute a valid formula. Such restrictions are not only logically inevitable; but if suitably chosen, lead to considerable economy in encoding the formulae and can also be used to give the system of formulae desirable mathematical properties, such as those of a lattice algebra mentioned above. In the interlingua I am now working on there are two such restrictions: (i) no term may appear more than once in a formula, and (ii) the terms are grouped in categories, such that no two terms of any one category can occur together. This leads to the set of all valid formulae forming a spindle product lattice (one of the types of canonical lattice) under the inclusion relation of sets of terms (being unstructured, each formula can be sufficiently represented as a mere set of terms).

IV. THE THESAURUS

Having got thus far, all we need to construct a interlingua on the required lines is (i) to decide on the minimal syntax needed; this we defer to the next section, and (ii) to select an appropriate repertory of

terms. This however is a very considerable task. An appropriate repertory of terms is virtually the same thing as a list of heads of a thesaurus. In a thesaurus, each word is defined and characterized by the set of heads under which it is entered; in our proposed interlingua, each formula will be characterized by the set of terms composing it. The mathematical correspondence is exact.

We have given a great deal of thought to the logical and mathematical structure of thesauri, and much of our ideas on this subject I shall not have time to go into here. A very detailed discussion of such matters will be found in a recent paper by Masterman (CLRU/ML90). I shall here confine myself to the more strictly mathematical description of a thesaurus.

The governing consideration in selecting our terms, and indeed all the thesaurus heads, however they are regarded, is that they should be demonstrably and reliably interlingual. For it is precisely as an interlingua, usable as an intermediate stage in the translation procedure between any pair of languages, that the thesaurus is useful. That is to say, our heads must convey ideas which are invariant under translation. In order to fulfil this requirement, an "idea" must be either (a) demonstrably common to all languages, or (b) dependent entirely on extra-linguistic facts. There is a limited range of grammatical functions which can be defined and identified in all languages, and which therefore qualify for inclusion under (a); it appears to be the case that the very sketchy grammatical analysis which these indications allow is nevertheless sufficient, with only slight help from rules peculiar to particular languages, to determine the structure of the input text: but for the present purpose this must be regarded as a fortunate accident. Also under (a) we can probably count the system of first, second, and third "persons," the numerals (at least up to ten), and probably nothing else. Under (b) however we can include a wide range of "ideas," divisible into (b1) specifiers of extra-linguistic contexts, and (b2) descriptors of material objects. It is true that not all material objects are known to all languages; but the absence in a given language of a word describing a foreign artefact, for instance, can be regarded as merely a gap in the dictionary, and does not invalidate the interlinguality of the system. Finally we may add (b3) descriptors of emotional states, subject to the provision that not all such states are as universal as some people think.

We have thus the following list of types of heads which we can make use of in a thesaurus:

1. Specifiers of interlingual grammatical functions.
2. Specifiers of number and person.
3. Specifiers of extra-linguistic contexts.
4. Descriptors of material objects and actions.
5. Descriptors of universal emotional experiences.

To which may be added, as common to all highly literate languages if not strictly interlingual: 6. Specifiers of style.

The problem now is, how many terms will be needed, in each of these classes, to provide a system capable of defining the meaning or range of meanings of every word in any language with sufficient precision to

make acceptable translation feasible. In other words, what is the relationship between number of terms and resolving power? Again, we can give no final answer; in fact, I can offer little more than conjectures. The difficulty is that it is almost impossible to assess the resolving power of any system without rather extensive tests. However, we can at least suggest some orders of magnitude, which will show that the system is not impracticable on this ground at least. Under (1), the number of functions which we can regard as interlingual is fairly small, certainly well under a hundred; and they are definable using not more than 10 descriptors, requiring 10 bits in the code. Under (2), we need to identify about 20 terms altogether, requiring not more than 5 bits. Under (3), we have generally worked on the assumption that what we need are roughly equivalent to the heads used in literary thesauri such as Roget's; if these are to be helped out by the other classes their numbers can be substantially reduced, to perhaps a few hundreds, and in any case the class can probably be encoded in not more than 40 bits. Under (4), the experience we have gained in the design of interlinguas suggests that something between 100 and 200 terms are required, mostly arranged in categories of 3 to 6 members; these may therefore need some 50 bits. The class of emotional descriptors is again fairly small, containing perhaps a score of terms, by no means freely combinable, and perhaps encodable in less than 10 bits. For style we need allow I think no more than 3 bits extra. This indicates that, I repeat as a conjecture, the requirement for adequate resolution may be of the order of 120 bits. If anything, this is likely to be an over-estimate.

V. INTERLINGUAL SYNTAX

As already explained, we shall aim to reduce to the minimum the amount of structural apparatus attached to the sequence of interlingual formulae which will represent the input text. It would be premature to be too dogmatic as to what this minimum is; all I can do is to outline the hypothesis which I am at present engaged on, and to contrast it briefly with other ideas being tested by the CLRU.

This hypothesis is that there is a one-to-one correspondence between the free clauses of any text and any acceptable translation of it in another language, and that within the free clause the whole of the relevant structure can be represented as a binary bracketing pattern with a single noncommutative and nonassociative combining operation between the formulae. It is clear that under certain circumstances, not yet fully determined, this combining operation behaves as if it were associative, so that a complete binary bracketing is not necessary; this weakens the structure still further.

The way this works out may be illustrated by a brief example. Here we take a simple English sentence; this is then rewritten, replacing those words which, I suspect, could not be rendered by single interlingual formulae, into pairs which could be so rendered; at the same time I introduce a bracketing pattern. In the last form, the words

are rearranged, and altered where appropriate, to illustrate the interlingual structure proposed: the single combining operation is easiest to understand if it is interpreted as a qualifier-qualified relation.

1. I went to the library to fetch a book
2. (This person) (((past go) (to (the (book storehouse))) (to ((carry back) (a book))))
3. ((This person) ((((a book) (back carrying)) purpose) (((the (book storehouse)) -wards) (past))) going)))) is

Notes: Since we are assuming one-one correspondence between free clauses it is necessary that these should be identifiable; the unit represented by "is" at the end is intended as a sign of a free clause; the form without this second moiety "is" would be understood as a nominal clause. Other units could replace "is" to indicate imperative and interrogative sentences. In each bracket group, we can interpret the first moiety as qualifying the second: thus, in "((the (book storehouse)) -wards)" the whole specifies a direction ("-wards"), and the form "(the (book storehouse))" specifies what direction is meant. The whole group in turn qualifies "(past going)" in the capacity of an adverbial group.

A few trials will show that any English sentence can be turned into this form with very little trouble. Rules quickly emerge; e.g. the form "subject-verb-object" regularly turns into "subject (object verb-noun)," and "preposition-noun" turns into "noun-verbal equivalent of preposition." An example of the latter would be "house occupying" for "in (the) house." The interesting point here is that precisely the same formal structures emerge even if we start from a quite different language. It must be understood that the words used in stage (3) of the above example would be represented by interlingual formulae of the kind already discussed; English words are used in the example only in order to allow the reader to identify the elements of the original sentence.

The example also shows one point where full binary bracketing can be dispensed with. This occurs (not necessarily only) where of a group of three units one is semantically much weaker than the others. Thus, in the form "(library -wards) go)" the form "-wards" is a weak one, possibly specifiable by a single term ("direction" will probably be included in our repertory); its presence therefore allows us to consider as an equally available alternative the bracketing "(library (-wards go))." In this form, the group "(-wards go)" could very well be rendered in proper English by some such word as "visit." Thus, we are able to pass from "go to the library" to "visit the library."

The very great weakening of syntactic structure which this type of interlingual syntax implies makes possible great flexibility in the build-up of syntactic forms at the output end of the procedure. If we now take English as the target language, let us see what we can do with the phrase which, with the formulae anglicized but the structure intact, could be written "((eye cheat) type) (hand skill)." The last group can be rendered, say, "dexterity"; the rest describes what type of dexterity. The formula represented by "type" could be rendered

by a relative pronoun, giving the possible translation "dexterity which cheats the eye." But this is only one possibility; for the weak element "type" indicates that reassociation of the first part is possible, giving "(eye (cheat type)) (hand skill)." The group "(cheat type)" would naturally yield in English a verbal adjective, i.e. participle, "cheating." Hence, we get the translation "eye-cheating dexterity." But this is not the end either; for it would not be hard to indicate that such forms as "eye-cheating" are stylistically somewhat specialized (the key entry would be "eye-" with its hyphen indicating object-relation), and if we were seeking for a more pompous style we could direct the rejection of "eye-cheating," and find (by application of the method of binary translation described later) some such alternative, involving a slight redistribution of heads between the two moieties, as "visually deceptive."

Previously, the CLRU has considered two other ideas of how to deal with syntax in translation, between which this is in some sense a compromise. We have done most work on the idea that one could ignore structure completely, and build up the whole syntactic structure at the output end from indications contained in special thesaurus heads selected with this end view ("syntax heads"). Though also highly flexible, this method imposes a heavy burden on the dictionary makers, who have to assign unambiguously and correctly a rather large number of not very closely defined syntax heads to every word. The other alternative goes to the opposite extreme of relying entirely on the bracket structure of the input text (elucidated with the help of word-class indications contained in the dictionary entries), which is carried over, wherever possible, unchanged; this method was originated by myself, and though admittedly too rigid, is easier on the dictionary maker, since he is required only to deal with a system of word-classes about equal in complexity to the traditional parts of speech system and so not too hard to learn.

VI. PROCEDURE

Let us now try to see how these various expedients fit together in the overall translation procedure. The whole procedure may be broken down into the following stages:

1. Read in	Takes written document and replaces each letter by a machine-readable code sign	Coded input
2. Look up	Identifies sequences of letters in coded input as dictionary headings in input dictionary, and replaces them by the readings given	Raw sequence

3. Pick out	Selects appropriate readings from alternatives offered, using unilingual rules; and eliminates noncontiguous bracket groups	Tidy sequence
4. Mark off	Identifies bracket groups among the formulae in tidy sequence, and reorders by syntax rules of interlingua; computes formulae for groups	Raw interlingua
5. Check over	Corrects the formulae in the light of context indications (including their own local interactions)	Tidy interlingua
6. Go across	Finds nearest equivalents in output dictionary for each formula, rejecting any not near enough, preferring large groups to small	Raw output
7. Turn round	Reorders the target language equivalents (in non-alphabetic code) by syntax rules of t.l., supplying particles, etc.	Tidy output
8. Write down	Looks up alphabetic code for each unit of tidy output and feeds to printing unit	Printed output

Of these stages, several are of no interest as regards the thesaurus and interlingual methods; we may thus ignore nos. 1, 2, 3, 7, and 8. Stages 4, 5, and 6 are of interest to us. The procedure for finding bracket groups is not what concerns us here; but it is an important part of the "mark-off" stage to compute a formula to represent, as accurately as possible, the semantic content of each bracket-group that is recognised. We have generally assumed that this would be done by taking the lattice meet of the formulae for the moieties of the bracket-group, but there is no reason why a more sophisticated operation should not be used, as it is unlikely that so simple a procedure will always work satisfactorily. The point is, that we should always give ourselves the chance of finding, in the output dictionary at stage 6, an acceptable rendering of quite large pieces of the input, at least up to the principal moieties of each free clause. This finding of the compound formulae for bracket-groups is one of the computations which make it necessary that the formulae on

which they are done should be firmly based in a workable algebraic system.

Further computations of the same kind are made at the "check-over" stage. The object here is to discover what heads (i.e. terms) are present in each formula which are repugnant to the locally relevant context, and delete them, thereby making the formulae a better indication of the relevant meanings. Thus, a "drive" is in English a thing which can be (a) walked along, and (b) taken part in (identifiable perhaps by the terms "route" and "activity"); in any particular context it should be possible to eliminate one or other of these terms by comparison with its neighbours in the bracket pattern. On the other hand, the retention of the "finance" meaning of the word "interest" would occur only if the general ("greater") context were biassed in this direction, as well as the local environment being suitable (even in a discussion on banking, one can show interest without reference to dividends).

VII, INEXACT MATCHING TECHNIQUE

However, the most interesting and important of the stages in which the thesaurus is used as such is no. 6 the "go-across" stage. It is here that the actual translation takes place. On entering this stage, the material will be in the "tidy interlingua" state, consisting of a sequence of interlingual formulae together with their relevant bracket structure and including a rendering, in the shape of an additional formula, for every bracket group of two or more formulae in the segment of text being processed. All these formulae will have been corrected in the light of the available context indications, so that each is the best we can do towards specifying what target language equivalent would satisfy us as a translation of the bracket-group to which it refers.

If possible, we shall try to translate a whole sentence or clause by a single word (rather, a single dictionary entry; there is no reason why the output dictionary should not include a certain number of quite long phrases if these cannot plausibly be built up from their component parts). In practice it is probably not worth trying this for a whole clause, because it will hardly ever succeed; but subject and predicate may both be put over in one word in many cases. There must be, of course, a criterion of whether we have succeeded in this. We shall presumably always accept an exact match between the given translation specification and what we find as the reading in the output dictionary; but this will be a rare occurrence, and we shall have to accept near misses as well. But we must not accept these too readily, or we shall obtain a very woolly translation. How to judge what is acceptable is likely to be a rather tough problem, but it is hardly likely to prove insoluble.

If a given bracket group yields no equivalent in the target language close enough to its specification to be acceptable, we must try next to

translate its two moieties separately. Note that this concept of "near enough" presupposes that the formulae and specifications are located in a metric space. If, as we intend, they constitute a lattice, there is no problem here; there are metrics available in plenty, and the choice of a suitable one is once more a matter ultimately for empirical testing. The criterion for translating the two moieties of a group untranslatable as a whole is the same as before; if either moiety has no equivalent near enough in the chosen metric to be acceptable, it in turn must be replaced by its moieties; and so on. However, it may well happen that we fail to translate a bracket-group consisting of only two formulae, and that we fail also to translate one of these by our chosen criterion. What do we do now? The answer is, that we must look for a translation, not in a single unit of the target language, but by a pair of units.

Thus, while we may reasonably give a universal preference to unitary translation, we must contemplate the possibility that in some cases binary translation will be needed. In binary translation we have to search in the output dictionary not for the nearest single entry to the given specification, but for the pair of entries, whose meet is nearest to the required specification (if, that is, it is indeed the meet that we use to form our compound formulae for bracket-groups). Algorithms for performing binary translation in this way can be constructed; though they are slow compared with unitary translation, they are likely to be practicable, provided they are not required too often. In principle we should also be prepared to try ternary and other multiple searches as well, but it is likely that these will indeed be unacceptably slow. In the end we must be prepared, as the human translator is too, though only on more severe provocation, to relax our criteria of an acceptable translation, rather than prolong the search for the mot juste uneconomically.

VIII. CONCLUSION

The preceding rather hasty sketch of the field will I hope serve to indicate some of the problems with which we are faced in devising a translation procedure using thesauric and interlingual methods. Perhaps it would be as well, before concluding, to mention some of the reasons which lead us to suppose that, in spite of the practical difficulties and theoretical complexities which this method entails, it is likely to be better worth pursuing than other methods which do not use a thesaurus.

First, it seems to me than none of the translation procedures so far proposed, other than those using a thesaurus, really face the problem of mathematicizing the procedure. If the whole procedure consists in dictionary look-ups of various kinds, the result can never be better than the dictionary makers provided for; the possibility that one can compute the best translation, given only a fragmentary knowledge of the possibilities embodied in separate words, is rejected. But it is

precisely in this possibility that the only hope of high-quality machine translation resides. For this purpose, some mathematically sophisticated system is imperative.

Second, it seems to me that the thesaurus principle, once it has been grasped (and it is not in itself so very abstruse) has an intuitive appeal which strongly suggests that it represents an important aspect of the functioning of language. We have already demonstrated the occasional power of the method to deliver high-quality translations; and the reasons for our frequent failures, when we were using Roget's thesaurus without emendations as our working model, were clearly traceable to defects in this document, rather than in the thesaurus principle itself. In the light of our experience to date, while we would be unwilling to guarantee that we shall ourselves solve all the problems now before us, we would be exceedingly surprised if there were an equally good solution for the Machine Translation problem not embodying the thesaurus principle in some form.

Third, we should not be frightened of the theoretical superstructure which our researches seem increasingly to require. If it is well built, such a superstructure is no hindrance to practice but a help to the design of improved methods. And if there were no accompanying theory, it is unlikely that whatever practical methods we evolved would have a more than temporary appeal even to the most practical minded, for the whole history of science and technology shows that rule of thumb methods eventually give place to methods based on more profound theoretical insights.