[From: *Computer Journal* 1 (3), 1958]

Interlingual Machine Translation

by R. H. Richens

Summary: The first part of this paper considers some of the reasons why mechanical translation via a logically formalized interlingua is worth pursuing. The interlingua described consists of a network of bonded semantic elements, the bonds being either homogeneous, corresponding to a generalized notion of qualification, or heterogeneous, for dyadic relations. The translation procedure involves a basic program applicable to any input language *(P)* and any output language (*Q*), and *P*-interlingua and interlingua-*Q* mechanical dictionaries. The essence of the program is the construction of an array of symbols, grammatical, syntactic and semantic, containing all the information required for translation. The interlingual translation of the input in *P* is then derived by successive eliminations, usually involving comparisons either across the rows of the array or down the columns. Similar treatment of a second array suffices to translate from the interlingua to the output *Q*.

The possibility of elaborating a completely general machine-translation program was first put forward in a paper read at the symposium on machine translation held at King's College, Cambridge, in August 1955, and published the following year (Richens, 1956, *a).* The next stage in this work was briefly reported at the Second Conference on Mechanical Translation at the Massachusetts Institute of Technology (Richens, 1956, *b),* since when considerable progress has been made towards reducing the procedure into manageable form.

INTERLINGUAL VERSUS DIRECT TRANSLATION

Machine translation via a logically formalized interlingua has called forth two opposing sets of criticisms. On the one hand, linguists have emphasized that two-step translation is invariably less accurate than one-step translation; on the other, machine translators have ventured the suggestion that an existing natural language could provide a satisfactory middle stage in a two-step translation program. One is hardly surprised to find that English and Russian are quoted as possible interlinguas.

Of these criticisms, the linguist's is the more cogent but applies only when natural languages are used as interlinguas. To translate, say, from Chinese to Japanese via English would be perverse. In proceeding from Chinese to English, for instance, the neutrality of Chinese words in respect of grammatical number has to be replaced by singular or plural determinations, and some attempt has to be made to introduce definite articles where English usage demands them. Japanese words, like Chinese, are often neutral with regard to grammatical number and seldom have anything corresponding to the definite article, although it is possible to achieve greater circumscription by using such devices as plural particles or demonstratives. Since English cannot without interpolated phrases represent the neutral situation, it is a poor means for connecting two languages where a neutral situation can be adequately symbolized.

There are three main reasons why translation via a formal interlingua is worth pursuing. Firstly, it seems to be the case that a rather crude level of mechanical translation between single pairs of languages is now in sight. The gulf between this level of achievement and that of a mediocre human translator is immense, however, and every step towards improvement is liable to make heavy extra demands on machine-dictionary requirements and sometimes also on the basic translation program. It is advisable, therefore, that whenever a linguistic problem is elucidated in respect of a particular language, whether serving as input or output in translation, it should be applicable to all circumstances, not merely to translation to or from one other particular language.

The second reason has been adumbrated already. Machine translation has a doubtful future with regard to such widely spoken languages as English, French and German; even Russian is unlikely to remain as little understood as at present. There is, however, never likely to be any general acquaintance with languages spoken by smaller groups, as for example Welsh, Albanian, Estonian, Georgian or Vietnamese, and if a Georgian speaker wishes to appreciate the imagery of Welsh poetry, machine translation might well provide an ideal approach.

The last reason is more theoretical. Linguistic and translation problems are, to one way of thinking, more clearly and usefully formulated in terms of a standard language, devised, as Wittgenstein (1922) once suggested, to mirror the logical multiplicity of the state of affairs which is being represented. Thus the twelve English terms *stallion, bull, ram, mare, cow, ewe, colt, calf, lamb, horse, ox, sheep* can obviously be replaced by three terms for the animal species and terms, respectively, for sex, masculine, youth and contrariety. It is redundant to allocate a term for female, which can be defined in terms of *sex, male* and *contrariety.* If preferred, feminists could define *male* in terms of *sex, female* and *contrariety,* but it is not possible to dispense with both *male* and *female.* Natural languages recede from formal simplicity in using homonyms. These have no place in an efficient interlingua.

THE INTERLINGUA

There are two principal aspects of any interlingual translation program that require consideration, firstly the nature of the interlingua itself and secondly the procedure for translating the input text into the interlingua and then translating the latter into the language of the output. In the line of research under discussion four interlinguas have been tried out; only the latest of these will be considered here.

This interlingua consists of a network of bonded semantic elements. Unlike the situation in natural languages, linear ordering has no significance, all syntactic relations being expressed by the linkages of the bonds. Two types of semantic elements are used. There are, in the first place, a limited number of primary elements, of the order of 50 to 100, representing such fundamental ideas as *exist, contrariety, cause, past in time, animal, perception, desire.* Many concepts can be completely defined in terms of these primary semantic elements. Thus *giving, receiving, donor, recipient, gift* can be defined in terms of the primary semantic elements denoting *cause* and *pertain.* When the primary semantic elements do not suffice for complete determination of a concept, use is made of arbitrarily numbered subcategories. Thus, the best equivalent that the primary semantic categories can provide for *canine* is a compound of the elements denoting *animal* and *pertain.* If, however, we recognize *dog* as a subcategory of animal and denote it, say, as *animal* 359, *canine* is completely definable. A primary semantic element may correspond to a word, as with *animal.* More often, it is a logical component of a word; thus *cause* can be regarded as a logical component of *give.* Less often, several words or word segments in combination correspond to one concept such as French *ne, pas* or the combination of noun suffix *-s* and the null verb suffix in English corresponding to *plural.* The primary semantic elements will be represented hereafter as lower case letters, though in the construction of mechanical dictionaries mnemonic catchwords are more useful.

The bonds linking the primary semantic elements are of two types, which may be termed homogeneous and heterogeneous. Homogeneous bonding corresponds to the usual idea of qualification. However, no distinction is made between the qualifier and the qualified; we do not distinguish between *black dog* and *canine blackness.* The relations between noun and adjective or between subject and intransitive verb are most frequently of this type. A homogeneous bond linking two semantic elements is represented by attaching the same superscript letter to each. Thus if b = *animal,* v = *male* and e = *emotional awareness,* $b^{a}v^{a}$ could represent a male animal while $b^{a}e^{a}$ could represent the statement that animals have feelings.

The heterogeneous bond is required for dyadic relations, exemplified by most prepositions and transitive verbs; the relation is distinguished by a superscript followed by 3, while the terms it connects carry the same superscript followed by 1 and 2. Thus if m = *mankind* and t = *cognition,* $m^{al}b^{a2}t^{a3}$ would represent somebody thinking about animals.

The order in which the semantic elements are bonded is indicated by ascending alphabetic order of the superscripts. Thus, if in the example quoted, we had someone considering asexual animals, this could be represented as $m^{cl}b^{bc2}s^{ab}c^{a}t^{c3}$, where s = *sex* and c = *contrariety.*

At the interlingual level there is no unit corresponding to the word; most natural words are semantically complex and so have a structure in the interlingua exactly like that between words. There are no distinctions between the elements in the interlingua corresponding to parts of speech. The only distinction is that some semantic elements require heterogeneous bonding while others do not.

INPUT→INTERLINGUA: GRAMMAR AND SYNTAX

So much for the interlingua. The procedure being developed for transforming input passages in a natural language into the interlingua will be outlined next. The objective of this work is complete generality, so that translation can be made from any language to any other without modification of the translation program. What are required, when translating from any language *P* to another language *Q,* are *P*-interlingua and interlingua-*Q* mechanical dictionaries. The *P*-interlingua and interlingua-*P* dictionaries are not interchangeable, but each should suffice for all translation from or to *P,* respectively. The dictionaries consist of several sections, the main being the list of word segments and the list of sequences of word classes.

Assuming that the input passage in *P* is scanned mechanically, the input operation consists in identifying or failing to identify the symbols, letters, punctuation marks and significant spaces, of as much of the input passage as the translation machine can absorb.

The essence of the program is the construction of an array of symbols, grammatical, syntactic and semantic, which contains all the information that may possibly be needed for translation. Redundant items of this information are then eliminated by successive operations, usually involving comparisons either across the rows of the array or down the columns.

The encoded input symbols form the first column of the array. Further columns of the array may be derived by comparing each input word with the *P*-interlingua dictionary, following, in essence, the procedure described at the First Conference on Mechanical Translation at the Massachusetts Institute of Technology (Richens and Booth, 1955). For each semantically significant segment of the input words, the following information is extracted from the mechanical dictionary: (1) word class, corresponding to a specially elaborated part of speech, (2) flexion class, such as first conjugation in Latin, (3) cross-reference data, for instance, referring the past tenses of strong verbs to their roots, (4) interaction with other word segments, and (5) interlingual equivalent. Obviously, one word segment may provide a number of rows of information, all but one of which will later be eliminated. Several cycles of dictionary comparison may be needed in highly inflected or compound-forming languages such as German or Finnish.

Cross-references also entail an extra dictionary comparison. It is, in addition, necessary to identify situations involving null flexions, such as the singular significance of the uninflected English noun, and add appropriate data for each to the array.

We have now reached a stage where grammatical analysis at a monolingual level can be accomplished. By means of comparisons down the word-class and flexion-class columns, problems involving flexion class can be resolved; thus *hoping* can be determined as pertaining to *hope* and not *hop.*

Syntactic monolingual analysis is principally achieved by comparison down the word-class column, matching word-class sequences against their occurrence in the *P*-interlingua dictionary, and generating a new word-class equivalent for the whole sequence. This procedure results in a new set of word classes, which is subjected to a second cycle of the same procedure, and so on until the entire syntax has been analysed. This type of syntactic analysis was first developed to deal with translation between particular pairs of languages. An example illustrating the application of the method to Chinese-English translation was described at the 1955 Cambridge symposium (Richens, 1956). A very similar method had been elaborated independently and published some months earlier in the United States (Yngve, 1955). In the present application, the method is used, not to obtain a rearrangement in word order, but to establish the correct bonding between the interlingual equivalents.

Other methods of analysing syntax have been proposed. In particular, Parker-Rhodes (1956) has devised an algorithmic procedure which is undoubtedly more mathematically elegant than that described here, but which seems to postulate rather more mathematical regularity in natural languages than they in fact exhibit.

Having analysed the syntax, it is possible to deal with such word interactions as have a syntactic component.

INPUT→INTERLINGUA: SEMANTICS

There remains the important residue of semantic problems that analysis of syntax, as commonly understood, cannot resolve. The human translator frequently decides that a particular rendering is “clear from the context.” This notion of clear inference from the context is remarkably obscure. In the program under discussion, three semantic analytic procedures are used, all based on comparisons down the interlingual column between the primary semantic elements of the interlingua. Firstly, there are semantic congruence relations. For example, the semantic element for cognition, t^{a3}, can be bonded to any x_1^{a2}, but it is necessary that $x_2^{a1} = m^{al}$, where, as before, *m* = *mankind.* This is merely a formal statement of the fact that anything can be thought of, but only human beings can think. Secondly, there are what might be called precise semantic determinations. Thus English *last* is likely to be an appliance only if shoe making is concerned, and not even then if the interaction *stick, to, last* is demonstrable. Thirdly, there are diffuse semantic determinations. If a word segment of ambiguous semantic content is compared with all the segments to which it is immediately bonded, and then, if necessary, to those more remotely connected, it is possible to make decisions as to its meaning based on the multiple occurrence of the component semantic elements. This sort of semantic comparison has been the object of extensive study by the Cambridge Language Research Group (see in particular Masterman, 1956), and forms one of the bases of the so-called thesaurus method. In the accounts of this procedure described hitherto, the sentence has formed the field of comparison, but this is certainly too wide in some instances. The method described above, in which comparisons are weighted in respect of bond connectivity, was designed to avoid the pitfalls involved in too wide a field of comparison.

It is possible that, having made a semantic analysis, all alternatives are eliminated. This indicates that the input words are being used more metaphorically than allowed for by the mechanical dictionary. It is necessary, in this case, to expunge the results of semantic analysis and to translate on the basis of syntax only. It is also possible that the syntactic analysis results in breakdown. This indicates either ungrammatical construction, which occurs too frequently to be disregarded in a machine-translation program, or stylistic innovation. In this case, syntactic analysis can be expunged and word-for-word translation provided.

The result of the stages outlined so far is a column of bonded interlingual semantic elements, which is the translation into the interlingua.

INTERLINGUA→OUTPUT

The following stages, translation into the language *Q* of the output, are analogous to those already described but with the following differences. Firstly, no word segmentation, cross-referencing, detection of null flexions or flexional analysis is necessary in respect of the interlingua. Secondly, in transforming the configuration of bonds into the syntax of the output language, re-ordering instructions are required since the interlingua is indifferent to word order and will merely reflect the original order of the input. It is, of course, necessary to construct the inflexions proper to *Q* if it is an inflected language.

Thirdly, while translation from *P* to the interlingua should be an exact logical transformation, translation from the interlingua to *Q* almost always involves some loss of precision, since almost all natural languages are incapable of representing some of the particular distinctions or particular vaguenesses occurring in other languages.

MECHANICAL ABSTRACTING

I have assumed that translation is what is required. It is possible, however, having achieved a translation into a logically formalized interlingua, to use it for other purposes. Attempts to devise mechanical abstracting procedures have tended to lean on statistical inferences, a very shaky approach, say, to a piece of tight argumentation. Some preliminary experiments suggest that it may be possible to use a formal interlingua as a basis for devising a mechanical abstracting program according to a logical specification. This possibility provides additional justification for the interlingual approach, but, since it lies beyond the terms of reference of this paper, it is a fitting point on which to close.

ACKNOWLEDGEMENT

Thanks are due to the National Science Foundation of the United States of America for supporting this investigation.

REFERENCES

MASTERMAN, M. (1956). "Potentialities of a Mechanical Thesaurus," *Mechan. Translation,* Vol. 3, p. 36 (Abst.).

PARKER-RHODES, A. F. (1956). "An Electronic Computer Program for Translating Chinese into English," *Mechan. Translation,* Vol. 3, p. 14.

RICHENS, R. H. (1956). "Preprogramming for Mechanical Translation," *Mechan. Translation,* Vol. 3, p. 20.

RICHENS, R. H. (1956). "A General Program for Mechanical Translation between Two Languages via an Algebraic Interlingua," *Mechan. Translation,* Vol. 3, p. 37 (Abst.).

RICHENS, R. H., and BOOTH, A. D. (1955). "Some Methods of Mechanized Translation," *Machine Translation of Languages,* edited by W. N. Locke and A. D. Booth, p. 24.

WITTGENSTEIN, L. (1922). *Tractatus logico-philosophicus,* p. 71.

YNGVE, V. H. (1955). "Syntax and the Problem of Multiple Meaning," *Machine Translation of Languages, loc. cit.,* p. 208.

DISCUSSION

Discussion following the presentation of Mr. Richen's paper at the Symposium on Mechanical Translation of Languages, held at Birkbeck College on 17 April 1958.

Mr. R. A. Fairthorne *(Royal Aircraft Establishment):* Associated with any set of languages there are two important interlinguae. One is the least refined language that makes all distinctions made by any of the original set of languages. The other is the most refined language that makes only those distinctions common to all the original set. This last, the *Pidgin* of the languages, is the best the translator can use without introducing information and principles not in the text and outside the scope of mechanical translation.

Neither interlingua need or should be suitable for human use. In particular, it can be synoptic, not serial.

'Language' is not confined to ethnic languages. Specialized and restricted languages for particular functions, commercial, scientific and administrative, are profitable fields for MT. Bibliographical search and 'information retrieval' in general are forms of translation, and are making very practical use of principles discovered in MT research, and conversely.

Mr. D. W. Davies *(National Physical Laboratory):* Mr. Richens described the symbolism in the interlingua for relations between two elements. Would he please explain how relations between three or more are symbolized? As an example, in the sentence 'John gave the book to the boy' the verb is a relation between John, the book, and the boy.

The author *(in reply).* I agree with Mr. Fairthorne's characterization of his two types of interlingua. However, the interlingua that I have in mind for machine translation is more general than either of these, since it was not derived by comparison between natural languages, and is intended to deal with new concepts that have not yet achieved symbolic expression in any natural language, or with concepts that have been symbolized in some other way. Natural languages are monolinear sequences of symbols; the interlingua that I have outlined is indifferent to order and arrangement of symbols, and can easily represent symbols arranged in branched or reticulate patterns.

With regard to Mr. Davies' point about the triadic relation *give,* this, like many other triadic relations, can be represented as two linked dyads. Thus, if x = the donor, y = the recipient, z = the gift, g = causation, and h = possession, we may interpret giving as causing somebody to possess something, and represent it as $X^{b1}y^{a1}z^{a2}h^{a3b2}g^{b3}$, the linked dyads $h^{a3b2}g^{b3}$ forming a triadic relation in respect of x, y, z.