SESSION 2

PAPER 5

TIGRIS AND EUPHRATES - A COMPARISON BETWEEN HUMAN AND MACHINE TRANSLATION

by

R. H. RICHENS

BIOGRAPHICAL NOTE

R. H. Richens was born in Penge, near London, in 1919. He read natural sciences at Cambridge and is now Assistant Director of the Commonwealth Bureau of Plant Breeding and Genetics at Cambridge. He has been a member of the Cambridge Language Research Group since its foundation. His principal research interests have been the taxonomy and history of the elm, the history of Soviet genetics, and machine translation.

TIGRIS AND EUPHRATES - A COMPARISON BETWEEN HUMAN AND MACHINE TRANSLATION

by

R. H. RICHENS

SUMMARY

TRANSLATION is treated as a species of formal transformation of symbols. Everything symbolized by a set of symbols constitutes the domain of symbolization of the set. The ultimate elements of the domain which symbolize nothing further are designated the terminal indicatum. Most domains of symbolization comprise mediate symbols which are both symbolized by other symbols and themselves indicate further symbols. Mental concepts are treated as symbols.

In translation, a set of symbols is transformed to another set in another language, the two sets having terminal indicata that only differ within narrow limits. Different kinds of translation can be differentiated on the basis of the extent to which the mediate symbols of the domains of symbolization of the input and output passages are similar.

Human translators utilize many categories of symbols and display great flexibility in their choice of procedure for translating and in their choice of criteria for ascertaining the significance of symbols of multiple use. Machine translation (MT) is much less flexible though most of the symbols used by human translators, with the exception of auditory symbols, phonemes and uttered-word segments, have analogues in MT.

Different types of input passage require different translation procedures, in particular with reference to the relative roles played by syntactic and semantic analysis. Closer formal resemblances may occur between human translation and MT procedures for the same type of input than between the procedures of either the human translation or MT confronted with input passages of various types.

While the over-all performance of the human translator is unlikely to be approached in the near future by MT, the latter may accomplish certain individual translation operations more efficiently than its human counterpart, and, on occasion, produce a better translation.

MACHINE translation (MT) promises to provide a highly developed analogue to an intricate operation till recently an exclusive pursuit of human beings. It is not the object of this paper to discuss this analogy either in the light of psychology or the philosophy of language; this can be done effectively only by specialists in these fields. What will be attempted here is to compare human translation and MT in as far as they are formal transformations, meaning by the latter any operation which converts one set of symbols into another. It will not prove possible to avoid all philosophic problems. Indeed, MT appears to be a breeding ground for new problems and new viewpoints concerning the philosophy of language. Such philosophic issues will only be touched upon here when they are inescapable.

Any passage in any language can be treated as an ordered set of symbols. By order is meant significantly arranged.

LANGUAGE AND SYMBOLISM

A symbol is anything, or a part or aspect of anything, or several things, or several parts or aspects of a thing or things which, either alone or in conjunction with other symbols, indicates something, usually but not necessarily, other than itself.

Curiously enough, there is no term in general use in English for whatever a symbol indicates. To facilitate discussion, the term indicatum will be used here. This is to be understood extremely widely; it may be one or many, it may be an object, a part, aspect, activity or state of an object, it may be a relation between objects, it may be anything that Wittenstein (1922) would have designated as a 'state of affairs', it may be a mental image or concept, it may be a falsehood, a contradiction or a tautology, it may be a symbol or set of symbols, and so on. The dyadic relation associating a symbol with its indicatum will be designated indifferently either by the verb 'indicate' or by the verb 'yield'.

In discussing what a particular set of symbols indicates, it is useful to establish its domain of symbolization. Thus a set of visual symbols, A, may indicate a corresponding set of concepts; these may indicate further concepts and these finally indicate a woman reading a book. The visual symbols, A, the two sets of concepts and the woman, the book and the act of reading together constitute the domain of symbolization. If the book is a novel it will contain printed symbols of its own but these will be excluded from the domain of symbolization of A. Similarly, A may be indicated by some other set of symbols B, but B is likewise excluded from the domain of symbolization of A.

In any particular domain of symbolization, a symbol which is not an indicatum of some other symbol or symbols will be termed an initial symbol. Similarly, an indicatum which is not itself a symbol yielding some further

indicatum will be designated a terminal indicatum. Symbols which both yield further symbols and are the indicatum of others will be termed mediate symbols.

The symbols of natural languages are of two principal types, auditory and visual, corresponding to spoken and written language respectively. It is convenient to treat the concepts corresponding to spoken or written language, or in fact any concepts, as symbols too since they also yield indicata. It is obviously not possible in a written paper to introduce other than visual symbols. We can however, introduce written symbols to indicate auditory symbols or concepts. The latter are of various categories, many being mediate symbols representing further concepts. It is convenient to distinguish between different categories of symbols by prefixing an indicator in the form of a capital letter followed by a full stop.

The following categories will be used:-

A	auditory symbol category
V	visible symbol category
P	phonemic category
G	graphemic category
U	uttered-word category
W	written-word category
L	lexico-grammatic category
S	syntactic category
N	'naked idea' category

All but the first two of these symbolic categories are concepts. When necessary, the categories can be subdivided by subscripts, e.g. L_1, L_2.

The significance of these categories is as follows. A.u: stands for a particular utterance of a long u: sound. It approximates to the [u:] of descriptive linguists. V.c stands for a particular token mark interpreted as of type c.

The A and V categories are initial symbols. A problem, however is involved in deciding what is to be regarded as an initial symbol. The letters of the roman alphabet seem to be fairly obvious initial symbols but should German V.ü, for instance, be regarded an an intial symbol, or an ordered arrangement of V.u and V.¨? When V.a and V.e are conjoined in the digraph V.æ, there would be some advantage in considering V.a and V.e as the initial symbols in English since variants with the letter disjoined will be frequently encountered. However, in Danish the digraph is regarded as a separate letter and might well be best regarded as the initial symbol. The Chinese ideographs present the problem in an acute form. If each ideograph be regarded as an initial symbol, many thousands would have to be recognized. On the other hand, it is possible to regard the ideographs as composed of ordered arrangements of a very considerably smaller number of units. Pushing the matter to an extreme, all written symbols could be

regarded as composed of a single visible symbol, replicated in two dimensions to correspond to whatever symbol is being considered. It is clear that the decision as to what constitutes an initial symbol must depend on the use to which this category is to be applied. A different decision might well be made for different natural languages.

P.u: stands for the phoneme u:, the universal concept to which pertain all sounds regarded as of the type u:, prescinding thus from pitch, intensity and other minor differences in pronunciation. In descriptive linguistics P.u: is commonly represented as /u:/. There are various interpretations of the phoneme concept among descriptive linguists; the one adopted here takes phonetic similarity as the criterion of phonemic identity, the phoneme being regarded as the referend of a phonetic type. However, it is desirable in the present context to extend the phoneme concept to cover all phonological features to which the concept of auditory type is applicable. Thus the tonic accent in Spanish subsumes a range of minor variations in pitch and intensity and can therefore be treated as a concept of the P category. G.c stands for the universal concept to which pertain all letters regarded as of type c. It thus subsumes handwritten c, typewritten c, and c in standard-roman and modern-face founts.

A question sometimes arises as to which symbols are to be regarded as of the same type. In a book set up entirely in italics, symbols in this fount might well be regarded as of the same type as the corresponding symbols in a book in roman fount. When, however, a few italic words occur in a passage otherwise in roman fount, it might be advisable to regard corresponding italic and roman letters as of different types since the italic words represent different indicata than the same words in roman fount. There are thus no formal criteria by means of which token symbols can be adjudicated to be of the same or different types. The decision must be made in the light of the application to which the type concept is being put.

Each symbol in the A category indicates a corresponding symbol in the P category and similarly for the V and G categories. Moreover, most graphemes in the G category also indicate phonemes in the P category since most written languages attempt to mirror speech. Cases do occur where graphemes indicate concepts but not phonemes, for instance the MT interlinguas of Richens (1956b, 1958), but usually, even in languages like Chinese where the ideographic role of the symbols in manifest, a phoneme sequence is also represented. Punctuation, however, may have no phonemic indication, as, for example the hyphen in the place name Weston-super-Mare. It also happens that a grapheme, say G.M, indicates a phoneme, P.m, and also a conceptual category, L.capital, with no corresponding phoneme.

Neither P.u nor G.c yields any particular terminal indicatum, or rather each potentially, that is, in combinations with other symbols, yields almost any indicatum. When combined with other symbols, P and G symbols yield relatively precise terminal indicata. Rarely, a single P or G symbol

does yield a relatively precise terminal indicatum as the Italian plural definite article P.i or the Polish preposition G.w.

An ordered combination of phonemes may constitute an uniflected uttered-word or a segment of an inflected uttered-word. Thus, the English phonemic sequence

P.d P.o P.g P.z

represents the uttered-word symbols

U.dog U.z

the former representing the animal and the latter plurality. Similarly, the graphemes

G.u G.n G.v G.e G.i G.l G.e G.d

represent the written-word symbols

W.un W.veil W.ed

It is probable that everybody has a concept of a word or word segment at a higher level of generality than the foregoing, that is prescinding from either the phonemic or graphemic categories. We may call this general category, the lexico-grammatic category. In descriptive linguistic studies, the term lexico-grammatical is sometimes used for words or word segments that partake of both lexical and grammatical characteristics as traditionally conceived; it is thus in the nature of a logical product. As used here, lexico-grammatic is far more equivalent to the logical sum of these two categories; it corresponds in a number of respects to the morpheme concept. The written-word symbols

W.un W.veil W.ed

would then indicate the lexico-grammatic symbols

L.un L.veil L.ed

The lexico-grammatic category requires subdivision, to allow for indication of the various types of grammatic category encountered. Thus French L_1.chien indicates L_2.masculine gender. The Latin affix L_1.itur indicates

L_2.third person L_2.singular L_2.present tense L_2.passive

The syntactic category comprises the indicata of the ordering relationships indicated either by single lexico-grammatic symbols or more usually by L symbols in some particular arrangement. Identical superscripts are used to indicate which symbols are bonded to which, bond signifying any syntactic connexion, however neutral.

Thus Latin

L_1.in L_1.flagrant L_1.e L_1.delict L_1.o

indicates

L_2.ablative adjective L_2.ablative noun

which indicates

S.qualifying adjective^a S.qualified noun^a

English

L_1.he L_1.cough L_1.ed

indicates

L_2.pronoun L_2.nominative L_2.verb ,

of which

S.pronoun subject^a S.verb predicate^a

is an indicatum; and Latin

L_1.carp L_1.e L_1.di L_1.em

indicates

L_2.verb L_2.accusative noun ,

which indicates

S.verb^a S.noun object^a

Lastly, for our penultimate indicatum, we have a category of 'naked ideas'. These made a somewhat provocative entry into MT discussions at the symposium on machine translation held at King's College, Cambridge, in August 1955. No one supposes, of course, that a concept can be discussed except by using language, but it is surely apparent that there are concepts that constitute a unique indicatum of numerous diverse renderings both within a single language and in different languages. Thus, there is a unique indicatum yielded by all the English words L.give, L.gift, L.present. L.donate, L.donation, L.receive, and Latin L.do, L.dono, L.praesto, L.largior, which symbolizes a state of affairs in which somebody A causes another person or perhaps an animal B to enter into a proprietary relation to an object C. Such a generalized 'naked idea' we can express as N.give. It is important to note that whereas any particular terminal indicatum may be symbolized in many ways using L symbols, only one set of N symbols is appropriate.

The N category may be subdivided, if the naked idea is analysable further. On this matter circumspection is necessary. Not only can numerous analytic schemata be devised for any one concept but the extent to which analysis is pursued in any one analytic scheme is frequently unlimited. It is also well to bear in mind Wittgenstein's (1953a) warning that an analysed statement does not necessarily constitute an increase in understanding over the unanalysed statement. The paradox of analysis, that analysis must be either trivial or false, is also a warning against analysis without a particular application in mind. Analysis in the present context is indeed trivial in as far as the final indicatum of any domain of symbolization remains unaffected. It is not trivial in as far as new mediate N symbols

are derived, which by virtue of their relations with one another, play a useful role in translation.

Thus, we have seen that L.give, L.present, L.donate all yield the same naked idea N_1.give. This, however is analysable further to

$$N_2.\text{cause} \quad N_2.\text{pertain}$$

with suitable syntactic bonding, where cause and pertain are catchwords for the corresponding general concepts. It is possible to analyse further, but in the present context it is doubtful if this would serve a useful purpose.

It is not implied that a human being, confronted with a linguistic passage, reacts by producing a domain of symbolization of the sort described or that all or any of the symbols set down above are indispensible in human translation. What does seem reasonable to assert in that each of the categories of symbols described are employed in human translation in some cases. It has been pointed out often that natural languages are over-determined and use several symbols to indicate a single point. Overdetermination, however, is usually a general characteristic and not a particular characteristic of any group of symbols so that, whereas it might be possible to ignore a symbol as redundant in one passage, it might well prove indispensible in another.

To exemplify a domain of symbolization, we will take the short English sentence, 'She came to'. The token symbols of this sentence are

$$\text{V.S} \quad \text{V.h} \quad \text{V.e} \quad \text{V.c} \quad \text{V.a} \quad \text{V.m} \quad \text{V.e} \quad \text{V.t} \quad \text{V.o} \quad \text{V..}$$

and the corresponding graphemes

$$G_1.\text{S} \quad G_1.\text{h} \quad G_1.\text{e} \quad G_1.\text{c} \quad G_1.\text{a} \quad G_1.\text{m} \quad G_1.\text{e} \quad G_1.\text{t} \quad G_1.\text{o} \quad G_1..$$

Of these, the first G_1.S indicates the corresponding lower case letter G_2.s together with a lexico-grammatic symbol L_3.capital.

The gramphemes indicate a series of phonemes which one may set down roughly as

$$\text{P.}\check{\text{s}} \quad \text{P.i:} \quad \text{P.k} \quad \text{P.e:} \quad \text{P.m} \quad \text{P.t} \quad \text{P.u}$$

and these yield the uttered-words

$$\text{U.}\check{\text{s}}\text{i:} \quad \text{U.keim} \quad \text{U.tu}$$

In addition to indicating phonemes, the graphemes also indicate the written-words

$$\text{W.she} \quad \text{W.came} \quad \text{W.to}$$

and the U and the W symbols both yield the lexico-grammatic symbols

$$L_1.\text{she} \quad L_1.\text{came} \quad L_1.\text{to}$$

The L category requires subdivision. Thus L_1.came indicates a verb the root form of which can be expressed as L_2.come; it also yields a grammatic

indicatum which may be designated L_3.past tense. The principal grammatic indicata in the L_3. category are

L_3.capital L_3.pronoun L_3.singular L_3.nominative L_3.verb L_3.past tense
L_3.verbal postposition L_3.stop

The L_3. symbols have indicata in the S category, expressible as

S.pronoun subject^{c} S.verb governing postposition^{abc}
S.verb qualifier^{b} S.verbal postposition^{a}

And lastly, we can proceed to the naked ideas

N_1.3rd person^{cd} N_1.female^{c} N_1.one^{c} N_1.become^{a}
N_1.conscious^{abd} N_1.past time-tense^{b}

which can be analysed one step further to

N_2.x^{d} N_2.female^{cde} N_2.one^{c} N_2.become^{a} N_2.conscious^{able}
N_2.past^{b3} N_2.communication situation^{b2}

The term x^{d} stands for the antecedent of the pronoun; past time-tense indicates past relative to the time of the communication.

The foregoing scheme provides one out of many possible bases for a formal description of translation which will be the topic under discussion in the next section.

TRANSLATION AS A FORMAL TRANSFORMATION

Symbols can be classified in many ways. For present purposes we can regard the corpus of symbols used by a group of people within which communication is easy as a language. Minor differences in usage within such a group can result in local dialects or technical expressions. Most languages are auditory and have a corresponding language in written symbols. Languages can also be classified in respect of subject matter; in some respects scientific Japanese is closer to scientific English than to literary Japanese.

A passage need not consist of a single language. Greek is a regular component of Hisperic Latin, English of medieval Law Latin and English and German of scientific Japanese.

A language need not consist exclusively of symbols. It may contain, for instance, sounds devoid of significance but with a certain quasi-musical role. Derry derry down, for instance, may convey a vague sense of joviality in some contexts, in others it signifies practically nothing.

Transformation of a set of symbols in one language to another set in the same language, with little change in the terminal indicatum, we may regard as constituting paraphrase. If the transformation is to a different language,

we may regard this as translation. It is rare for a translation involving natural languages to yield exactly the same indicatum as the original passage since most languages have particular shades of meaning or particular vaguenesses that are not simply expressible in other languages. The amount of difference in the indicata of the original passage and its translation, and the nature of the differences in the two indicata that are tolerated in translation are a matter of choice. They might well differ according to the particular purpose for which the translation is being made.

Translation, moreover, may involve further considerations. The domain of symbolization of a particular set of initial symbols is usually highly elaborate. It may be required to effect a translation in which, not only the terminal indicatum, but various of the mediate symbols are either kept unchanged or altered to only a slight degree. Thus, whereas the terminal indicatum of a Latin sentence in the passive voice, i.e. with the symbol L_3.passive in its domain of symbolization, can be translated into an English sentence in the active voice without change in the terminal indicatum, it is often required to conserve such a symbol as L_3.passive during translation. Similarly the Italian noun

$$L_1.\text{prod} \quad L_1.\text{uzion} \quad L_1.\text{e}$$

can be translated into English as an abstract noun,

$$L_1.\text{produc} \quad L_1.\text{tion}$$

or as a gerund,

$$L_1.\text{produc} \quad L_1.\text{ing}$$

The terminal indicatum is unaffected but the former translation involves fewer changes in the mediate L_3. symbols in the domains of symbolization.

It is implied in the previous paragraph that a basis for comparison exists between different language in respect of their mediate symbols. Considerable progress has been made towards an exact description of the A, V, P, G, V, W, L and S categories of symbols within a single language; this in fact is the subject matter of descriptive linguistics. The techniques of investigation of descriptive linguistics, however, lean heavily on commutability relations which are practically nonexistent between different languages. Comparative grammar has been investigated but no sufficiently exact system for establishing formal comparisons between grammatic symbols in different languages has yet been developed. Allen (1953) has shown along what lines descriptive linguistics might be expected to contribute to formalized comparative grammar; Lambek (1958) has approached the problem from the mathematical angle.

Thus, there is obviously a close relationship between the English mediate symbol L_3.noun and the Italian L_3.noun. However, the former can be frequently used as an uninflected qualifier preceding another noun while the latter cannot. There are perhaps no exact grammatic equivalences

between different languages and, at the moment, we are hampered by lack of suitable techniques for characterizing what partial equivalences there are with sufficient precision.

Some types of translation aim at conserving the auditory symbols as far as possible, that is the symbols of category P. Verse translation is a related exercise; in this case it is not the phonemes that are conserved but a phonemic pattern, or what would more usually be described as a phonological pattern which, as the indicatum of the P (P_1) symbols, we may designate as category P_2. The preservation of alphabetic patterns in psalm translations and the preservation of acrostics are likewise instances in which P_2 symbols are as far as possible kept invariant.

It is possible to translate according to more complex criteria, for instance onomatopeic proportionality, in which highly onomatopeic words in the original passage are rendered by onomatopeic words in the translation and conversely for words with little onomatopeic suggestiveness. This procedure conserves particular relation between the P and N symbols. It is not necessary here to pursue such refinements further.

From what has been said above about translation, it would appear that the process of analytic understanding is itself a form of translation since it, like translation, involves transformation of symbols with invariance of the terminal indicatum. Each of the categories of symbols enumerated can in fact be regarded as a language on its own, so that, in as far as translation between natural languages involves mediate symbols, it can be regarded as a concatentation of subordinate translations. The notion that analysis is a form of translation has been discussed from the philosophic angle by Langford (1942).

A final topic that requires discussion before human translation and MT can be compared in detail is the range of use of symbols. In any particular passage, each symbol signifies something and is therefore precise in some sense (cf. Wittgenstein, 1953b). In the sentence 'My cat likes rum in her milk', V.cat refers precisely to a particular species of animal. There are however numerous varieties of cats and in respect of varietal distinctions, the sentence is vague. All ambiguous passages are vague at one level of precision, yet if a passage represents either N.x or N.y, it is precise if taken as representing N.x or y.

A great many symbols yield very different indicata in different contexts. In the sentence 'My sister let the cat out of the bag', the indicatum of V.cat is vastly different from what it was in the earlier example.

A distinction is frequently made between normal and metaphorical usage of a symbol. It is not clear whether a formal distinction between these two usages can be established. It is certainly the case, however, that most symbols have a range of use and that the determination as to which use they have in a particular passage is a crucial operation in both human translation and MT.

The decision as to which use a symbol has in a particular passage depends on its relations with other symbols. Two principal types of relevant relationship can be distinguished, (1) relations between lexico-grammatic and syntactic symbols in the L and S categories and (2) relations between naked ideas in the N category. These two types of relation are not of course to be considered as exclusive. To exemplify, the coexistence of the symbols L.let, L.cat, L.out, L.bag affords a strong presumption that the unwitting release of information is being indicated. In the Italian word sequence

$$L_1.la \quad L_1.prod \quad L_1.uzion \quad L_1.e$$

L_1.la indicates L_3.definite article; in other contexts it indicates a preverbal pronoun but this usage is excluded here since

$$L_1.prod \quad L_1.uzion \quad L_1.e$$

indicates L_3.noun and not L_3.verb.

The significance of 'plant' in the English passage 'a pineapple slicing plant' is revealed by a combination of relationships involving both S and N categories. This passage has the following indicatum

$$L_3.indefinite\ article \quad L_3.noun \quad L_3.verb\ stem$$

$$L_3.present\ participle \quad L_3.noun$$

which represents

$$S.noun\ object \quad S.verb \quad S.noun\ agent$$

Then, finally, the decision as to whether L_1.plant represents the botanical entity, N_1.plant, or a machine installation, N_1.machine, can be approached by examining the congruence between the indicatum of S.verb, i.e. N_1.slice, and the two potential indicata of S.noun agent, i.e. N_1.plant and N_1.machine. N_1.machine and N_1.slice are fully congruent but N_1.plant and N_1.machine are less so and therefore N_1.machine will be chosen as the N_1.indicatum of L_1.plant.

This preamble provides one means of examining the translation process as a whole and we can now pass on to a survey of the various operations involved and of the different ways in which these are performed in human translation and in MT.

HUMAN TRANSLATION AND MT

The first translation operation is recognition of the passage to be translated. In human translation, this involves either auditory or visual perception. The allocation of symbols to phonemic or graphemic categories is accomplished with great facility by the human translator, though it is almost invariably the case that recognition is partly based on context. This is shown by the frequency with which printer's errors are overlooked

in proof, the reader recognizing a letter as otherwise than what it is owing to the context in which it occurs. In reading handwriting, a great deal of recognition depends on context.

Till now MT has principally proceeded from written texts and the initial recognition and subsequent encoding has been done by a human operator. Photoelectric scanning is, however, feasible even though it is hardly yet a practical alternative, but is not likely to reduce all the variant token symbols of a single type to their G indicatum anywhere near so readily as the human translator. It may, for instance, be necessary to use separate matching operation to identify, say, roman, italic or gothic founts representing the same letter. Then, by encoding all variants of the same grapheme in the same way, reduction to the same G symbol can be accomplished.

Any photoelectric scanning procedure will probably have to allow for a margin of uncertainty in identification. How much deviation can be tolerated is a matter which the operator must decide. As mentioned earlier, a choice has to be made as to what is to be regarded as an initial symbol. This need not be the same in human translation and MT.

It is probable that the human translator is seldom unaware of the phonemes and uttered-words represented by a written text. In cases when script is vaguer than the spoken language, as in the use of invalid in 'His grandfather left an invalid will', and 'His grandfather had been an invalid for years', the human translator will probably elaborate different P and U symbols in the domains of symbolization of the two passages.

In MT it is possible to establish auditory distinctions of the sort mentioned but there is no point in doing so since the distinction can be treated as a particular instance of range of use. Current MT procedures, therefore, differ from human translation in having no analogues for A, P and U symbols.

Having recognized the symbols of the input, it is possible both in human translation and MT to proceed unanalytically. In the case of the human translator, this could involve no more than matching the entire input against a phrase book and writing out the equivalent provided. No understanding of either input or output is required; the only requirement is ability to derive the graphemes of the input symbols. Given a German-Italian phrase book, an English operator, ignorant of either foreign language, could correctly translate from German to Italian whenever the phrase book included the set of symbols of the input.

An analogous unanlytic procedure is feasible mechanically and once an input passage has been recognized and encoded, the problems involved in matching it against a mechanical phrase book, finding the equivalent, and putting this out, are trivial. However, it is obvious that such a procedure is of practical value in only a very limited number of cases. It may have some relevance in highly stereotyped situations such as those involving business correspondence or ordering meals in foreign restaurants.

Already, however, we are faced with the pervasive problem of dictionary searching. The human translator, in searching for equivalents, may rely on his memory, an extremely efficient operation from the point of view of access speed but less efficient from the point of view of comprehensiveness of entries and freedom from errors. Alternatively, he may consult a dictionary, usually alphabetically arranged; this is a fatiguing procedure, though the storage capacity of a printed dictionary is high.

Computer storage media are being so rapidly developed that it would be premature to state what the practical limits of storage capacity and access time are. To date, three principal dictionary searching procedures have been considered for MT: direct alphabetic searching, reverse alphabetic searching (Richens and Booth, 1955) and the partitioning or bracketing method (Booth *et al.*, 1958). The first and third are suitable for invariant phases or uninflected words since decomposition of the phrase or word into smaller units is not required. Under these circumstances, the method of Booth *et al.* is the speedier. The reverse alphabetic method is probably best for dealing with languages containing inflected words, since this method permits both decomposition of inflected words into their constituent units and location of the units in the mechanical dictionary. This method, however, involves an analytic operation in addition to recognition and matching of the input symbols and is more properly discussed below.

Phrase books normally connect two languages directly but it is possible to conceive of phrase books in which translation is achieved in two stages via some interlingua. One set of phrase books would be of the form, input language - interlingua, and the other, interlingua - output language. Perhaps the only advantage of such a procedure in unanalytic translation would be the reduction in the number of phrase books required for translation in all directions between n languages: $n(n-1)$ in the case of ordinary phrase books and $2n$ in the case of the interlingual method. It must be remembered that neither natural languages nor the current artificial languages can serve as satisfactory interlinguas since it is seldom the case that a translation into any of these has exactly the same indicatum as the passage translated. It would also be possible for unanalytic MT to proceed via an interlingua though this would be of little practical interest either. The sole advantage of unanalytic interlingual translation is the reduction in total dictionary requirements. When, however, analytic MT translation is being considered, several additional advantages of interlingual methods become manifest and under these circumstances they became a practical issue of some importance.

In analytic translation, the input passage is no longer treated as a single unit but as a set of symbols which can be analysed.

In both human translation and MT, the input symbols are analysed into ordered sets, each yielding a word or word segment as its indicatum.

The human translator may utilize uttered-word segments (U symbols), written-word segments (W symbols), or generalized lexico-grammatic symbols of the L category, but for MT only the L category is normally considered. It is possible to treat entire words as the units for memory or dictionary consultation but it is more usual in human translation and of greater general utility in MT for the search unit to be a word segment. Thus L_1.dogs will not be found in a normal dictionary but L_1.dog will. In MT, mechanical dictionaries have been devised that take the entire word as the search unit. Such dictionaries make heavy demands on storage space and lengthen the access time required for a single search through the entire dictionary, particularly in highly inflected or compounding languages such as German, Finnish or Hungarian. A very considerable economy in storage requirements is effected by using a mechanical dictionary of word segments, stems and affixes, each of which has a significance or a range of significance of its own.

We are confronted, then, with the necessity of deciding which symbols, phonemes or graphemes are to be associated together in the words or word segments. Most modern written languages indicate the boundaries of words by spacing. In Arabic script, additional indication of the word boundaries is provided by the special forms of initial and final letters. In Japanese, however, the boundaries between words in hiragana script are not indicated and have to be ascertained by other means.

The boundaries of word segments are more difficult to discover and the human translator employs many criteria, probably discerning the most distinctive segment first and then identifying the rest by elimination. Thus in

G.d G.o G.g G.s

he will first isolate L_1.dog; in

G.j G.o G.k G.i G.n G.g

L_1.jok; and in German

G.g G.e G.t G.r G.a G.g G.e G.n

L_1.trag. The human translator tends, after isolating a word segment such as L_1.jok, to transform this into a conventional form such as L_1.joke which is an actual word representing the segment and is used by dictionary compilers in arranging their entries. This transformation of word segments to a conventional word is redundant in MT, which can operate directly with L_1.jok, which is entered as such in the mechanical dictionary.

The human translator can pick out distinctive word segments from the beginning, interior or end of words. This generalized operation is not easy to mechanize and the usual procedure is to identify word segment from the beginning of the word. One of the most effective methods (Richens and Booth, 1955) is to compare words against a mechanical dictionary arranged

in reverse alphabetic order and to effect a match when the word is first wholly contained within the dictionary entry. This procedure would result in misdivision of the word in some cases, thus discontent is liable to be divided into disc-on-tent. This problem has been discussed by Richens and Halliday (1957); the general solution for the difficulty is to place awkward words as separate undivided entries in the mechanical dictionary.

The purpose of recourse to memory or to a dictionary or grammar is to obtain as many indicata in the domain of symbolization of each L symbol as are requisite for translation, and also to obtain appropriate corresponding symbols in the language of the output. Whenever a human translator utilizes symbols of the N category, he is in fact translating interlingually, since as noted above, only one set of N symbols is appropriate to any particular terminal indicatum. It is probably exceptional for human translation to proceed without any recourse to symbols of this degree of generality.

Most MT procedures have been devised to proceed from a particular input language to a particular output language without recourse to an analogue of the N symbols. In favour of this procedure is the simpler programme. Richens (1958) has urged the advantages of interlingual MT, on the grounds, mentioned above, that fewer mechanical dictionaries are required, and has pointed out that even if an analogue of the N symbols is dispensed with in obtaining equivalent symbols from the mechanical dictionary, no method of solving the semantic problems connected with multiple use without recourse to N-symbol analogues has been devised. The type of interlingua envisaged is a logically formalized standard language fulfilling, as far as possible, Wittgenstein's (1922) criterion that a linguistic passage and the state of affairs symbolized by it should have the same logical multiplicity.

Between direct MT translation, a one-one procedure, and Richens' interlingual method, a many-many procedure, there are the many-one methods of Masterman (1956, 1957), Halliday (1956) and Parker-Rhodes (1956b). In these, translation proceeds from any input language to the output language via a mechanical thesaurus in the output language. This approach is referred to in more detail below.

Returning to the example quoted earlier, this would be encoded mechanically in some such form as

$$L_3.\text{capital G.s G.h G.e G.c G.a G.m G.e G.t G.o G..}$$

and this, utilizing the ordering information provided by the linear sequence of the symbols and by spacing, can be matched against a mechanical dictionary to yield the sequence

$$L_1.\text{she } L_1.\text{came } L_1.\text{to}$$

Since none of these words is inflected, problems attending decomposition into word segments do not arise. In languages where a single word may consist of up to five segments, not uncommon in German or Finnish, these problems may be of some complexity.

The problem of multiple use arises as soon as dictionary matching has been achieved. Thus, if we had the English word L_1. plant, it might indicate the botanical entity, a machine installation, the act of inserting the botanical entity into the soil, the act of founding a community, or the perpetration of a fraud.

The human translator's approach to the problem of deciding the use intended is one of extreme flexibility, and many different criteria may be used for the various cases within a single sentence. The tendency in MT is to solve these problems as early in the programme as possible since the solution of one such problem is often a prerequisite for the solution of another.

Thus, in such a passage as 'He planted a new colony', corresponding to

$$L_1.\text{he}\ L_1.\text{plant}\ L_1.\text{ed}\ L_1.\text{a}\ L_1.\text{new}\ L_1.\text{colony}$$

the fact that L_1.plant is followed by L_1.ed suffices to eliminate the uses, botanical entity or machine installation, but is of no help in deciding between the remaining uses.

In the case of

$$L_1.\text{she}\ L_1.\text{came}\ L_1.\text{to}$$

the problem of range of use will probably be deferred in MT till L_1.came has been collated with the dictionary entry L_2.com. It is possible to deal with the range of use of the various forms of strong verbs independently of each other but this would entail quite unnecessary duplication of dictionary material for each entry. It is probable too that human translators consider L_1.came and L_1.com as members of a single more general category whose range of use can be considered as applying to all its variants.

The MT procedures just described differentiate between inflexion and ablaut. Thus from

$$G.c\ G.a\ G.m\ G.e$$

we proceeded to L_1.came and thence to L_2.com.

$$G.k\ G.i\ G.c\ G.k\ G.e\ G.d$$

however, would yield

$$L_1.\text{kick}\ L_1.\text{ed}$$

with Latin

$$G.r\ G.e\ G.x\ G.i\ G.t$$

would yield

$$L_1.\text{rex}\ L_1.\text{it}$$

and L_1.rex would be referred to L_2.reg.

The distinction should not be pressed. L_1.sang could be made to yield

L_2.s-ng L_2.-i-

and quite possibly this sort of symbolic relationship is utilized by human translators.

In MT, it is easier to handle affixation and ablaut separately and therefore the structure of the domain of symbolization of the two is differentiated. Whether as L_1.came or L_2.com, we are confronted with a wide range of use. L_1.com, considered alone, can indicate motion towards, or happen, and in combination with other L_1. symbols we have the uses indicated by come about, come across, come by, come home to, come in for, come into the world, come of, come on, come to a head, and come to, the latter combination indicating either motion towards or recovery of consciousness.

It is doubtful whether either in human translation or MT, a decision as to the use of L_2.come or L_2.to can be decided without reference to the relations between L_3. symbols which yield the S categories. L_1.she has a narrower range of use and, if a nautical context can be excluded, will seldom be other than the third person singular nominative female pronoun. The L_3.symbol sequence has been given as

L_3.capital L_3.pronoun L_3.singular L_3.nominative L_3.verb

L_3.past tense L_3.verbal postposition L_3.stop

The human translator will have little difficulty in inferring that the L_3.capital ... L_3.stop sequence bounds the syntactic unit termed the sentence. This can also be ascertained mechanically though it is necessary to discover whether the capital preceded a word normally capitalized and whether the final stop is a sentence-conclusion indicator, an abbreviation indicator or a combined abbreviation and sentence-conclusion indicator.

The nominative pronoun in an initial position will suggest to the human translator that it is the grammatic subject of the sentence, the following verb will be taken as the predicate and the final particle with no following noun or noun equivalent will be attached as a postposition to the verb. It is important to note that the syntactic relations are described here as derived from ordered L_3. grammatic categories. This is not invariably so and the flexibility of the human translator is shown again in the way in which other criteria, such as semantic congruence, can be used in establishing syntactic relations.

For instance, in the Japanese sentence 'ruku ka sen shoku tai ni zigzag jo nomonogo mii da sareta', i.e. 'zigzag configurations were found in the hexavalent chromosomes', the postposition L_1.ni is translated an 'in'. It is also used however as the agent of passive verbs. The verb is passive here but L_1.ni is not taken as the agent since chromosomes do not make observations.

MT syntactic analytic methods for deriving the S categories are far more stereotyped than those of the human translator and have tended to be based solely on the relationships between grammatic categories. On the other hand, though the MT methods for syntactical analysis are sterotyped, they have sometimes been based on principles that have no obvious parallel in human translation.

Among the MT syntactic analytic techniques with close analogies to those of the human translator are those of Yngve(1955) and Richens(1956a). In these, syntax relations are determined by detection of sequences of grammatic classes of word segments, the significance of the sequence being obtained by consulting a mechanical dictionary of such sequences, in effect a mechanical grammar. Parker-Rhodes (1956a) has experimented with a technique of syntactic analysis which is algorithmic and aims at performing syntactic analysis with no or little recourse to a dictionary of grammatic forms. The efficiency of this method is not yet known, but should it prove effective, it would provide a good example of an MT procedure based on an algorithm with no obvious analogue in human translation.

In human translation and most types of MT, when the syntactical patterns in the S categories have been determined, an appropriate arrangement of symbols in the output language can be obtained by reordering transformations and the provision of syntactic flexional forms. Thus, from Spanish 'una casa grande' reordering 123 $\rightarrow$ 132 yields English 'a large house'. English 'the eagle does not catch flies', does not require reordering to obtain the Latin 'aquila non capit muscas', but needs appropriate affixes for the grammatic subject and object of the verb.

In the interlingual MT technique of Richens, syntax is expressed solely in terms of a configuation of syntactic bonds at this stage and is left as such in the interlingua without rearrangement. Reordering is only introduced when proceeding from the interlingua to the output language.

Once the syntactic relations within a passage are determined, the range of use of the symbols narrows greatly. Thus, in our example, the facts that there is a syntactic bond between L_1.came and L_1.to and L_1.to is not followed by a noun or noun equivalent are sufficient, both in human translation and MT to indicate that

$$L_1.\text{came}\ L_1.\text{to}$$

is a semantic unit and to exclude the signification 'motion towards'.

Syntactic analysis resolves many points that arise during translation, but there remains a residuum of problems concerning range of use that depend for their solution on semantic analysis. Here, the powers of the human translator are seen at their subtlest, the range of criteria used in semantic analysis being extremely wide. In one of the examples quoted in the previous section, the derivation of the significance of the phrase 'letting the cat out of the bag' would probably be made on the basis of collocation of symbols in the L category. An example where the use was

determined by semantic congruence of N symbols was also given and it is easy to think of cases in which the use of a symbol is decided by some item of knowledge, either scientific or historic. Thus, in the sentence 'Jane Shore was the mistress of Edward IV', the decision as to the significance of 'mistress' might well be made on purely historical grounds.

The mechanization of semantic analysis is perhaps the most recondite of MT problems, especially as it is probably the case that not one but several types of semantic analysis are required for adequate translation.

Booth *et al.* (1958) deal with semantic problems by characterizing each use of a symbol as pertaining to one or other of a limited number of semantic fields. The field is either assumed to be known in advance or is determined during the course of translation by counting unambiguous words.

A far more elaborate development of the semantic field conception has been made by Masterman (1956, 1957 and personal communication), Halliday (1956) and Parker-Rhodes (1956b). In their approach, the thesaurus method, semantic analysis proceeds concurrently with the treatment of syntax. The first operation in this method involves obtaining for each L symbol, a series of thesaurus heads, corresponding roughly to N symbols, covering the entire range of use of each L symbol. Thesaurus heads represented once only in a sentence are eliminated by diallel comparisons of the sets of heads for each L symbol. Next, L symbols with relatively invariant significance are withdrawn from further comparison. Then, finally, a splay of synonyms in the output language is obtained for each remaining thesaurus head, and the synonym or synonyms common to the remaining heads of each L symbol are used to provide the basis for the output.

The semantic analytic technique of Richens (1958) differs from the foregoing in being more tightly associated with syntactic analysis which it follows. Three separate semantic analytic operations are envisaged: (1) semantic congruence between syntactically bonded N symbols, (2) identification of special collocations of L symbols, and (3) a form of the semantic field method.

It is interesting to recall at this point the Japanese sentence considered above. It is doubtful whether syntactic analysis together with semantic-field considerations could correctly determine the use of L_1.ni in this sentence. However, the semantic incongruence between the N indicata of the verb and of the putative agent is sufficient to determine the alternative use of L_1.ni and hence to derive the correct syntactic bonding.

A difficulty arises at this point. Neither human translation nor MT is feasible without some semantic analysis yet passages are frequently encountered where the metaphorical use of words presents serious difficulties. Take Crashaw's

> ..... soft powers
> Whose silken flatteries swell a few fond hours
> Into a false eternity.

A pedestrian semantic analytic programme may splutter at powers being soft, at flattery composed of silk, at hours being either fond or swelling and at eternity being false. There is every possibility that semantic incongruences will be detected in such a passage and it is necessary, when these would otherwise inhibit translation, to relax too tight a form of semantic analysis or even discard it entirely and translate on the basis of syntactic coherence or semantic-field considerations.

In the example we have been following through,

$$L_1.\text{she}\ L_1.\text{came}\ L_1.\text{to}$$

yielded

$$L_3.\text{capital}\ L_3.\text{pronoun}\ L_3.\text{singular}\ L_3.\text{nominative}\ L_3.\text{verb}$$
$$L_3.\text{past tense}\ L_3.\text{verbal postposition}\ L_3.\text{stop}$$

whence we obtained

$$\text{S.pronoun subject}^{c}\ \text{S.verb governing postposition}^{abc}\ \text{S.verb qualifier}^{b}$$
$$\text{S.verbal postposition}$$

$L_1.$come, when bonded to $L_1.$to and in the absence of a noun or noun equivalent governed by $L_1.$to, forms a semantic unit with $L_1.$to which would transform the sentence into the form.

$$\text{S.pronoun subject}^{a}\ \text{S.verb predicate}^{a}$$

If a MT programme for direct translation from English to French is being run, reordering would be accomplished at this point. In this instance the reordering is of the form 1 2 → 1 2 and no change is required. However, assuming that the verb is rendered by the French 'se remettre', the correct reflexive pronoun has to be introduced between the subject and the verb, and assuming that the perfect tense is used, agreement has to be established between the gender of the subject and the past participle. The final output would be 'elle s'est remise'. Should an interlingual MT programme be used, an interlingual rendering of some such form as

$$N_2.x^{d}\ N_2.\text{female}^{cde}\ N_2.\text{one}^{c}\ N_2.\text{become}^{a}\ N_2.\text{conscious}^{able}$$
$$N_2.\text{past}^{b3}\ N_2.\text{communication situation}^{b2}$$

described in the first section, will be derived. It is possible at this point to regulate the type of translation according to the degree of similarity in the mediate symbols required. If the grammatic categories of the input are to be mirrored in the output, we might obtain such renderings as French 'elle s'est remise', Latin 'animum collegit', or Japanese 'sho ki ninatta'.

It is relevant to point out that none of the translations proposed, nor in fact any likely to be offered by a human translator, have the same terminal indicatum as the input passage. The French verbal stem is vaguer

but its tense, with its perfective aspect, more precise, the Latin fails to indicate the sex of the subject and the Japanese, literally 'there was a becoming to consciousness', indicates neither sex of the subject nor differentiates between first or third person or singular or plural number.

In direct MT, the shift in the terminal indicatum may take place at any stage in the programme. In interlingual MT, no shift in the terminal indicatum should occur till after the interlingual rendering has been derived.

A few remarks should perhaps be added on some rather special cases, which though of infrequent occurrence cannot be totally ignored in any formal study of translation.

It has already been noted that semantic analytic procedures encounter difficulties with some types of metaphor. Similar difficulties may arise in passages containing logical contradictions or contradicting definitions of words in the mechanical dictionary. An extremely rigorous semantic analytic programme might even throw out tautologies as vacuous. Further it would be unwise to depend too exclusively on syntax since sentences defective in grammer or syntax are not infrequently encountered.

Sentences such as 'pig has three letters', where the terminal indicatum of 'pig' is a symbol and not the animal also require care, lest, for example a translation such as French 'porc a trois lettres', be obtained. It would be feasible in such cases, though perhaps not worth while, to include special directions in the mechanical dictionaries concerning entries referring to symbols, so as to avoid this trouble.

In such sentences as 'Chic is a foreign word', foreign is relative to the language of discourse and so, in translating into French or any other foreign language, is best rendered as 'foreign to English'. There is also a convention that passages in quotes should be translated, even though their function is often to indicate the original form of a communication.

These examples are not quoted as serious problems, but merely to stress that neither human translation nor MT are procedures to which a single rule of transformation is applicable. Human translation is characterized by its flexibility and the varied methods it employs for dealing with different types of passage. The need for differentiating between situations in which syntactic analysis must be paramount and those in which semantic analysis is of prime necessity has already become evident in MT, and it is possible that the power of discriminating between procedures rather than the intensive development of the procedures themselves will be the major MT research problem of the future. There is no immediate prospect of MT rivalling the all-round performance of the human translator though it is highly probable that individual translation operations might, in some cases, be more efficiently accomplished mechanically, with, on occasion, the production of a translation superior to human.

ACKNOWLEDGEMENTS

The preparation of this paper has been greatly assisted by discussion with Professor R. B. Braithwaite, and Miss M. M. Masterman, Dr. M. A. K. Halliday and the other members of the Cambridge Language Research Group. Thanks are due to the National Science Foundation of the United States for a grant in support of this work.

REFERENCES

1. ALLEN, W. S. Relationship in comparative linguistics. *Trans. Philol. Soc.*, 1953, 52.
2. BOOTH, A. D., BRANDWOOD, L. and CLEAVE, J. P. *Mechanical resolution of linguistic problems*, 1958.
3. HALLIDAY, M. A. K. 1956. Linguistic basis of the thesaurus-type mechanical dictionary and its application to English-preposition classification. *Mechan. Translation*, 2, 36. (Abst.)
4. LAMBEK, J. The mathematics of sentence structure. *Amer. Math. Monthly*, 1958, **65**, 154.
5. LANGFORD, C. H. The nation of analysis in Moore's philosophy. *The Philosophy of G. E. Moore*. edit. P. A. Schilpp. 1942.
6. MASTERMAN, M. Potentialities of a mechanical thesaurus. *Mechan. Translation*, 1956, **3**, 36. (Abst.)
7. MASTERMAN, M. M. The thesaurus in syntax and semantics. *Mechan. Translation*, 1957, **4**, 35.
8. PARKER-RHODES, A. F. An electronic computer programme for translating Chinese into English. *Mechan. Translation*, 1956a, **3**, 14.
9. PARKER-RHODES, A. F. Mechanical translation program utilizing an interlingual thesaurus. *Mechan. Translation*, 1956b, **3**, 36.
10. RICHENS, R. H. Preprogramming for mechanical translation. *Mechan. Translation*, 1956, **3**, 20.
11. RICHENS, R. H. A general programme for mechanical translation between any two languages via an algebraic interlingua. *Mechan. Translation*, 1956b, **3**, 37. (Abst.)
12. RICHENS, R. H. Interlingual machine translation. *Computer J.* 1958, **1**, 144.
13. RICHENS, R. H. and BOOTH, A. D. Some methods of mechanized translation. *Machine Translation of Language*, 1955, edited by W. N. Locke and A. D. Booth, p.24.
14. RICHENS, R. H. and HALLIDAY, M. A. K. Word decomposition for machine translation. *Rep. Eighth Annu. Round Table Meet. Linguist. Language Stud., Georgetown, Washington, 1957*, p.79.
15. WITTGENSTEIN, L. *Tractatus logico-philosophicus*, 1922, p.71.
16. WITTGENSTEIN, L. *Philosophical investigations*, 1953a, pp.21-24.
17. WITTGENSTEIN, L. *Ibid.* 1953b, p.45, p.139.
18. YNGVE, V. H. Syntax and the problem of multiple meaning. *Machine Translation of Language, loc. cit.*, 1955, p.208.

DISCUSSION ON THE PAPER BY MR. R. H. RICHENS

PROF. Y. BAR-HILLEL: Mr. Richens, as most of you probably know, is one of the pioneers of machine translation. A paper he wrote, in collaboration with Booth, in 1948, which was reproduced in 1952 for the *First International Conference on Machine Translation* and finally published in 1955 *(ref. 1)* was the first serious contribution to machine translation altogether, and I am sure that everybody who was working on MT in these early times drew great profit from studying it. Unfortunately, however, during the last years Mr. Richens went off into lines of thinking which I do not believe will be very fruitful. I have no time to discuss here the issue of interlingua which was the major topic of Mr. Richens' oral presentation - I do this, however, in a forthcoming report on "The present state of machine translation in the United States and Great Britain" -. I shall rather go into the theoretical foundations which Mr. Richens provides us with in his paper. I personally am much more interested in these foundations than in the practical applications and was therefore highly disturbed by the way Mr. Richens accomplishes this job.

Mr. Richens' approach relies heavily on two terms, 'indicate' and 'concept'. Both terms are notoriously vague in ordinary usage and one might have expected an attempt of clarification on behalf of one who is going to use them so extensively. However, what is meant by 'concept' is never discussed at all. With regard to 'indicate' and its counterpart 'indicatum', not only is no indication of their prospected usage given, but it turns out that Mr. Richens uses these terms such that a sound indicates a letter, a letter a word, a word a concept, a concept other concepts as well as other entities. Why Mr. Richens should believe that by employing a single term for a whole gamut of what are usually regarded to be rather different relations he will gain something, is beyond me. I am sure that nothing useful can result from such a treatment.

Mr. Richens' pair of terms, 'indicate' and 'indicatum', is obviously coined to replace the more customary terms, 'designate' and 'designatum', 'denote' and 'denotatum', 'signify' and 'significatum'. Mr. Richens' preference might be due to the fact that the other terms were not always used with sufficient rigor. It may also be due to the fact that some of these terms were used, e.g. by Rudolf Carnap in his books *(refs. 2* and *3)* with too high a degree of rigour for his purposes. At any rate, some indication of the relationship would have been helpful.

Refs. 1, 2 and *3* on page 305.

There are quite a few additional passages which must cause a great deal of confusion. Take the definition of 'symbol'. A symbol, according to Mr. Richens, page 282, is anything which either alone or in conjunction with other symbols indicates something. From this definition I gather that a symbol which indicates something *only* when in conjunction with other symbols does not indicate anything by itself. The insight that not every symbol symbolizes is as old as Aristotle. 'Syncategorematic' was the term used for this kind of symbols by the Schoolmen. Unfortunately, Mr. Richens himself looses very soon sight of his own definition and talks as if every symbol had its indicatum.

There is also a lot of confusion in the use of the term 'initial symbol'. You find on page 282 that an initial symbol is a symbol which is not an indicatum of some other symbol or symbols. However, a page or two later the same term is used as a synonym for 'simple symbol' in contrast to 'complex symbol' rather than to 'mediate or terminal symbol'. As a result of this confusion, you will find on page 283 the problem raised whether the German 'ü' should be regarded as an initial symbol or as an ordered arrangement of 'u' and Umlaut. I do not think that this question has anything to do with the question whether 'ü' is an initial symbol in the sense defined on the page before. The issue is, of course, whether 'ü' should be regarded as a simple or as a complex symbol, which is a totally different distinction from the initial-noninitial one.

On page 286, the term 'naked ideas' is introduced and we are told of its provocative entry in 1955. It is well known, of course also to the Cambridge Group itself, that similar conceptions played a big role in medieval speculations as well as in the speculations of such authors as Descates, Leibniz, Locke and Bishop Wilkins in the 17th century, and that they did not do much good. Though this fact does not prove that a similar idea could not meet with better success in the 20th century, no real argument is given why one should expect a change. At any rate, I do not understand what it could possibly mean that 'give', 'gift', 'present', 'donate' and 'donation' have a unique indicatum. Without a more or less complete system telling in detail what difference it makes whether one assumes that these words do or do not have a unique indicatum, I cannot see how anything of consequence can be said about the subject. A couple of sentences are surely not enough to endow it with any scientific import whatsoever.

Let me conclude with a few minor comments.

I was disturbed to find on page 289 a reference to a paper by Lambek to the effect that Lambek has approached the problem of formalising comparative grammar from a mathematical angle. I know this paper pretty well. What Lambek does there is to develop further, especially in the algebraic direction, a notation for syntactic description which I proposed a few years ago *(ref. 4)*, which itself was based on certain ideas brought up by the Polish

Ref.4 on page 305.

logician Ajdukiewicz *(ref. 5)*. I cannot see what this has to do with comparative grammar, though I can see its connection with what has been termed 'universal grammar'.

I am also profoundly disturbed by the fact that Prof. Braithwaite is mentioned as one of the people who assisted Mr. Richens in discussion of his ideas. I would be highly gratified to learn that Prof. Braithwaite is not at all to be held responsible for any of the ideas I have been criticising. I can hardly believe that he would entertain such ideas, knowing Prof. Braithwaite personally.

In the last part of his paper, Mr. Richens makes some very interesting remarks on certain difficulties in machine translation which might have been noticed before but were never discussed before in print, perhaps because there were graver problems that required discussion. However, I would not agree with the way in which Mr. Richens almost dismisses the problems he pointed out himself. He mentions, for instance, the problem of translating into French such a sentence as "Pig has three letters" (where 'pig' is written without quotation marks, as even scientists - especially in England, I am told - are in a habit of doing). Translating this sentence as '*Porc a trois lettres*' is clearly wrong, as already indicated by the fact that the English sentence is true and the French one false. However, "*Jean a quatre lettres*" is a wrong translation of "John has four letters" though both sentences are true. I do not believe at all, in distinction from Mr. Richens, that it would be easy to program a machine to come with the correct translation "'*John*' *a quatre lettres*". Even more generally, I am convinced that the treatment of machine translation of material in quotation marks (or of material that should have preferably been printed in quotation marks) will cause a lot of trouble, to such a degree, indeed, that I regard it as highly unlikely that a fully automatic treatment of such material will at all be possible. By the way, I know that human translators have troubles with such material, too. But this does not mean that the problem is not serious or that its treatment can be postponed for long.

Another problem is the translation of such a sentence as "Chic is a foreign word". Translating this into French as "Chic [or 'chic', for that matter] est une parole étrangère" is plainly wrong since 'chic' is not a foreign word (in French). 'Foreign' is a context-dependent word, and

REFERENCES

1. RICHENS, R. H. Some methods of mechanized translation in Machine Translation of Language ed. by W. N. LOCKE and A. D. BOOTH (1955).
2. CARNAP, R. Introduction to Semantics. *Harvard University Press*, (1942).
3. CARNAP, R. Meaning and Necessity. *University of Chicago Press*, (1947).
4. BAR-HILLEL, PROF. Y. A quasi-arithmetical notation for syntactic description. *Language*, 1953, 29, 47.
5. AJDUKIEWICZ, K. Die syntaktische Konnexitat. *Studia Philosophica*, 1935, I, 1.

translating an English sentence containing this word into another language changes the context decisively. An unsophisticated translation of such a sentence will therefore not be a good translation.

Again, contrary to Mr. Richens' opinion, I believe that the problem involved is serious. There is no simple procedure to find out which, and in what way, the words of the English language are context-dependent. And I don't think that the issue can be belittled for the reason that context-dependent words do not occur in scientific discussions and writings. They might not be too abundant in ordinary scientific papers on matters physical or chemical, but there would surely be plenty of them in discussions of matters linguistic, for instance. This might be one reason why so far hardly anybody has tried to machine translate papers in linguistics. As soon as this is attempted, the seriousness of the problem will become immediately evident.

MR. R. H. RICHENS (in reply): I appreciate Prof. Bar-Hillel's stimulating comments on what he takes to be my philosophical views. I should however point out that my paper was not primarily intended as a contribution to philosophy nor am I a philosopher. The general principles set down in my paper were not developed first and then applied to mechanical translation; rather, the machine translation schedules were devised first, with a great deal of trial and error, and then afterwards it proved possible to discern certain principles of general application.

Prof. Bar-Hillel states that he would be highly gratified to learn that Prof. Braithwaite were not responsible for the views that he has criticized. I can offer Prof. Bar-Hillel little solace here. Prof. Braithwaite is, of course, in no way responsible for any of the views set out in my paper, but he is cognizant of the entire conceptual scheme developed, he was present at the research meetings of the Cambridge Language Research Group at which these views were exhaustively discussed, he made a number of suggestions that were incorporated into the final text, and, I am assured, he is not in strong disagreement over any major issue. The term 'indicatum', to which Prof. Bar-Hillel takes a strong dislike, was adopted after a discussion of Prof. Bar-Hillel's alternatives with Prof. Braithwaite. Each of these alternatives carries implications which I would repudiate in the present context.

Prof. Bar-Hillel fails to understand why the term 'indicate' is used to cover 'a whole gamut' of relations. I think that, on this point, he has confused vagueness with generality. I am well aware that the connections between the various categories of symbols that I have used are diverse. However, I imply nowhere that 'indicate' is a monolithic notion. What I do assert is that there are features common to all the categories of 'indicate' which justify the use of a single term. Moreover, there is a precise

analogue to 'indicate' in the machine-translation programme on which I have been working, namely substitution of a term or terms by the following entries in a mechanical dictionary or inventory of higher-order categories.

The reproof that I talk as if symbols indicate in isolation is rebutted by my definition of a symbol. I did not include examples of indication by pairs or multiple associations of symbols in my paper; they are common in practice and duly allowed for in machine-translation programmes.

Finally, about naked ideas. I know that they have been floating around the world for a very long time. Whether or not they have been misused by others is not my concern. They are adopted here for a precise practical end, the resolution of ambiguities requiring semantic analysis. Possibly Prof. Bar-Hillel can see nothing in common in the words 'give', 'receive' and Latin 'do', 'dono' and 'praesto'. Most people, excluding philosophers and linguists, can. Even if they could not, the naked ideas discernible by semantic analysis are still useful in mechanical translation for the purpose I have mentioned.

After all, the justification for the principles I have enunciated, assuming that they have any, is not that they accord with any particular philosophical outlook but that they are effective in producing specimen translations from and to diverse languages. I assert nothing more.