ABSTRACT.

THE RELEVANCE OF LINGUISTICS TO MECHANICAL TRANSLATION

by

Martin Kay

Linguistics is relevant only to certain restricted and clearly-defined aspects of Mechanical Translation. In particular, the Immediate-Constituent model is relevant to input routines, and the Transformational Model is relevant to output routines. Since these do not cover the most important part of the translation process, a middle stage has to be inserted, even when translating between only two languages. It is convenient to call this an Interlingua since the information from the input stage is here expressed in a wholly formal calculus. From this fact, it follows that this middle-stage vehicle of information must be an artificial construct and not a natural language.

The case made here for the insertion of an Interlingua is not the normal one, since the main thesis of this paper is that high-quality translation is incompatible with preserving the dichotomy between grammar and lexicon. The Interlingua must constitute an area in which syntactic and semantic information from the input language can be reallotted as between the syntactic and lexical units of the output language.

The use of such an Artificial Interlingua is also justified on three other important but less fundamental arguments:

1. *The number of routines will be proportional to the number of languages rather than to their square.*
2. *The compilation of dictionaries, etc. for esoteric languages will be made more practicable.*
3. *An artificial Interlingua could be made to introduce less noise than a natural language used as an Interlingua.*

These constitute the normal justification for Interlingual Mechanical Translation. But the linguistic lacuna which requires an Interlingua manifests itself even before the question of Interlingual Translation comes up.

THE RELEVANCE OF LINGUISTICS TO MECHANICAL TRANSLATION

by
Martin Kay
June, 1959

It is one of the requirements of the "scientific method" that each part of each application of it must be made explicit [1]. The situation is rapidly being reached in some sciences where the languages we know, aided as they are by other symbolic systems, no longer seem to be equal to this task. It is already the case that some apparently cognate sciences and branches of sciences use mutually unintelligible languages. In these circumstances, it is hardly surprising that there is growing concern that our knowledge of the workings of language should be extended as much as possible. One obvious approach is to attempt to apply these very scientific methods to language itself[2]. It is also not surprising that such an attempt entails linguistic problems peculiar to itself, for in this study language is at the same time the subject of the investigation and an essential part of the apparatus for performing the investigation. This difficulty, which philosophers have always recognised, is often treated in too cavalier a manner by linguists. H.J. Uldall[3] says:

> "It remains to examine one more hindrance, viz, the curious fact that the humanities so to speak contain themselves; form part of their own material."

- but then goes on to dismiss this hindrance on the ground that there is no subject of which this cannot be claimed. Inasmuch as our brains are made up of elements known to chemistry, we are using a chemical tool to operate on the subject of chemistry and therefore chemistry, in Uldall's sense, contains itself. However, the logical process involved in building up an abstract system are not a direct function of chemistry in the same way as they are of language. A scientific system is not itself a chemical compound;

---

1. "Whatever may be the case with less precise forms of thought, deductive thinking is not independent of the possibility of its expression." R.B. Braithwaite, Scientific Explanation, p. 24.
2. "One can only hope that linguists will become increasingly aware of the significance of their subject in the general field of science". E. Sapir, "The Status of Linguistics as a Science", Lg. 5 (1929).
3. "Outline of Glossematics" (p.2), Travaux du Cercle Linguistique de Copenhague, vol.X1. (1957)

it is, however, a linguistic construct. Chemistry belongs to the "real" world; the world we live in is the world of language.[1]

At first sight, then, the new scientific study we call "Linguistics" may seem to have contributed only to the chaos of a situation it might have been expected to clarify. Linguists have, in fact, distinguished themselves no more than other scientific workers by the way they have kept their terminology within bounds. They have split into factions, each with its own terminology, and in such a way that the division between the two main groups corresponds to three thousand miles of water.

It is inevitable that in the context of Mechanical Translation, we shall be giving the larger part of our attention to the activity to which the term Linguistics is normally applied in America, for it is this which aims at a more specifically mechanical description. Allusion to the European approaches will be made later. My aim will be to substantiate the thesis that Mechanical Translation requires a new approach to language which will not be founded on present-day Linguistics. I shall try to show that any approach which divides language into semantic and grammatical aspects, and studies one or the other, or both, separately, will not provide answers to the problems of Mechanical Translation. I shall try to make a case for a method which regards these two aspects as essentially complementary so that the shortcomings of the techniques applied to each may be supplied by the other.

The tendency of the American school of "Descriptive" or "Structural" linguists has been to stress the importance of their empirical approach and the necessity of embodying their results in a deductive system. They have taken upon themselves the task of describing language without reference to historical or pragmatic considerations. Einar Haugen[2] has the following to say about this as opposed to what he

---

1 An extensive treatment of this problem is included in "What is a Thesaurus?" by M. M. Masterman (in this volume),

2. "Directions in Modern Linguistics", Lg 27, (1951).

calls the "traditional" approach to Linguistics:

> "It will be my thesis that any linguistic entity can be described from two points of view, one internal to the language described and one external to it; further that traditional linguistics has sought objectivity by adopting an external standard to which the language may be referred, while present-day linguistics seeks to find internal, relational standards, and finally, that while the internal or distributional standards may lead to useful discoveries concerning the internal organisation or structure of the language, linguistics cannot, unless it wishes to become entirely circular or mathematical, afford to reject the use of external standards to give its relational data concrete validity in the real world".

Given the explanation of this use of the word "mathematical" which Einar Haugen provides, this is also my thesis. One of the consequences of using language as a tool to investigate language is that it is unusually difficult to ensure that external, and in particular, intuitional standards are not in fact being applied. Few linguists claim that they have succeeded in totally excluding such standards, but they claim that they are invoked only in the interpretation of formal linguistic systems in terms of particular languages, and not in the formal systems as such. Martin Joos likens these systems to maps and points out that no logical means are available for showing that a map represents a given piece of ground. This, however, no more prevents explorers using maps than it does physicists using mathematics. He says:

> "The place for logic is inside the map, not between the map and the real world."(1)

My purpose is to show that, in the context of Mechanical Translation at least, this no longer holds true.
The map analogy is powerful. It is also used by H.J. Uldall [(2)]:

> "An autonomous discipline can be built up only by resignation, by being willing to do one thing at a time, by a rigid selection of a set of functions as necessary and sufficient for unambiguous description, i.e. by abstraction. You cannot make a map if you insist on bringing in all the hills, valleys, houses and trees in life size and complete to the last wood-louse."

---

1. Martin Joos, "Description of Language Design", Journal of the Acoustical Society of America, 22 (1950).
2. loc. cit.

A map, in fact, must be designed to fulfill a very specific purpose. Either it must be a road map, or a geological map, or a map showing underground electrical cables or something of the kind. The questions we have to ask about Linguistics as contributing to research in Mechanical Translation are simply:

1. Is the map to the right scale? Could we manage with a smaller one, or must we ask the linguists to try and give us a more detailed one?
2. Do we need a different kind of map or do we want another kind of information added to the one we are given? If the linguist gives us all the roads we need, do we want the relief as well?
3. If we need a different map, will the one the linguist supplies help us to make it? If we want to know where the drains are, and we know they only run under roads, a road map will be of some use.

Let us now turn our attention more closely to "Descriptive Linguistics" with the demands of Mechanical Translation specifically in mind. In what follows, the words "Mechanical Translation" are to be understood as meaning "Fully-Automatic High-Quality Mechanical Translation". To many, and in particular to the inventor of this phrase[1], this will seem to betoken an unwarranted idealism on my part. I justify it simply as follows. In the first place, considering the problem in this context will, I believe, enable us to speculate more fruitfully about possible future developments both in Linguistics and in the field of Mechanical Translation. Low-quality mechanical translation may be achieved in the more or less distant future in a variety of ways. However, if the ideal of High-Quality mechanical translation is realised, it is likely to be as a result of a small number, probably only

(1) Y. Bar-Hillel, "Report on the State of Machine Translation in the United States and Great Britain, February 1959.

one, of these ways having proved intrinsically superior to the rest. If we are to speculate, as we must in a field where so much lies in the future, let it be about this method. Secondly, I am unconvinced that the unfeasibility of Fully-Automatic High-Quality Mechanical Translation has been demonstrated. I have not the courage to fall in with those who are prepared to write off such an endeavour as hopeless at so early a stage.

The aim of the linguist is to discover and describe those features of a language which characterise it as a vehicle for information. This is his criterion of relevance. Now, by causing someone to speak into a microphone and by operating in various ways, and with various devices on the resulting electrical signal, it is possible to arrive at interesting results about the sounds which were made. The microphone and associated devices themselves suggest analytical techniques and, together with these techniques, set limits on the features of the subject matter which can be investigated and the scope of the final description. The analysis will, in all probability, be in terms of harmonic analysis or some other essentially quantitative technique. In fact, the limits which this sets on the analysis are such as to include little of interest to the linguist. The linguist's first task, then, is to find the tools and techniques most appropriate to his criterion of relevance.

> "First we must limit our field, leaving outside it certain things to be treated precisely by engineers or sociologists, while we speak of them more or less artistically. Second, within our field, we must adopt a technique of precise treatment which is by definition a mathematics. We must make our linguistics a kind of mathematics within which inconsistency is by definition impossible" [(1)].

This is a very strong definition, so strong, in fact, that there are many linguists, even in America, who would prefer not to limit themselves so closely[(2)]. However, the more Linguistics moves in this direction, the more it can be expected to serve our more specifically mechanical purposes.

1. Martin Joos, "Description of Language Design", Journal of the Acoustical Society of America, 22 (1950).
2. "In view of the fact that methods as mathematical as the one proposed here have not yet become accepted in linguistics, some apology is due..." Z.S.Harris, "From Morpheme to Utterance", Lg.22 (1950).

To the credit of American Linguistics must be ascribed the devising of a body of objective techniques which limit the field within the criterion of relevance, though it remains for us to see whether the limits set are not, in fact, too close for our purposes.

Now, there are in general two types of mathematical systems which a scientist may call upon. There is that which would be appropriate to the kind of analysis we have mentioned, using electronic apparatus to examine the sounds of speech, a type which uses the Infinitesimal Calculus, Fourier Analysis and the like. It is the type of mathematics which is applied to subjects which can be envisaged as systems of continuous variables. The second is an essentially discrete kind of mathematics which has been applied recently in a number of new fields, notably in Quantum Mechanics. It is a system of this latter type which structural linguists have chosen to use.

> "All continuity, all possibilities of infinitesimal gradation, are shoved outside linguistics in one direction or the other. There are, in fact, two such directions in which we can and resolutely do expel continuity; semantics and phonetics" [1].

One of the first and most interesting findings of Structural Linguistics" was that there are in the study of any language, clearly delimitable areas in which a finite, discrete model can be made to yield significant results.

Clearly, Linguistics, like any other science, must pass through a "natural-history" stage in which the basic work of classification is done and on which any system which will eventually give unity to the subject, will be based. It is more than usually difficult in the study of language to decide when this natural history can be said to have reached an advanced enough stage to provide the basis of a formalised system. The very rigorous limits which the linguist sets upon his activities require that he should apply himself only to extant texts, and there is no reliable way of knowing how much text constitutes a valid sample of any language. There are a great many important phenomena in every language which occur in extremely weak dilution. However, for the natural history alone, for a

---

1. Martin Joos, loc.cit. I am very much indebted to this article for helping to give shape to the ideas in the first part of this paper.

system of methods and units, we must be truly grateful to Structural Linguistics. Furthermore, a number of linguists have already made interesting excursions into the next phase and others have declared this to be the object of their endeavours[1].

We are justified, therefore, in including as part of the ultimate aim of Linguistics the attempt to identify such features of a language as can be mapped[2] onto a formal system, that is, to construct a calculus which can be interpreted in terms of a natural language. Similarly, Mechanical-Translation research seeks to identify the set of features of a language which, being mapped onto a corresponding set of features in another language, represent the maximum information content[3]. These characterisations, I believe to be accurate. They give weight to the view that Linguistics and Mechanical-Translation research are, so to speak, made for one another, and that any advance in one will be, ipso facto, an advance in the other. I think we shall find this is misleading.

I shall assume for the moment that, setting aside as it does, certain features of its subject to which the discrete model of its choice seems inappropriate, there are in language empirically discernable features to which this model does correspond. These features, I shall, for

---

1. "Nærvaerende sprogteori har fra sin første planlægelse været inspiriret af denne erkendelse, og tilsigter at tilvejbringe en saadan immanent sprogets algebra", Louis Hjelmslev, "Omkring Sprogteoriens Grundlægelse", p. 72.

   "Essentially linguistics is a sort of logical calculus, although the analogy must not be pushed too far", J.B. Carrol, "The Study of Language".

   "Hence, the investigation of a language entails not only the empirical discovery of what are its irreducible elements and their relative occurrence, but also the mathematical search for a simple set of ordered statements that will express the empirical facts", Z. S. Harris, "Distributional Structures", Word, 10 (1954).

   "The object of linguistic analysis is the establishment of the minimal arbitrary code on the basis of which the facts of speech may be understood", C. E. Bazell, Word 10 (1954).

2. "Mapped" is here used in the informal sense given to it on page 3.

2. No specific reference is here made to the "Information Theory". The word is used with its non-technical meaning.

simplicity, call "formal". Thus the distinction between noun and verb, or active and passive is "formal", whereas a distinction made on semantic grounds alone is not.

Now, the hypothesis that Mechanical Translation is possible implies that utterances can be found in one natural language which are, in some sense, more or less perfect maps of utterances in another. It implies a criterion according to which an utterance in one language can be said to be more or less similar to an utterance in another. An intuitive criterion we know to exist, for without this there could not even be human translation, however imperfect. If we have ever done any translation, we further know that this criterion is not formal in the meaning we have given to the word. It is not required of a good translation that it should exhibit a "formal" resemblance to the original. It thus becomes apparent that, whatever the contribution of Linguistics may be, translation holds problems which lie without its terms of reference and, indeed, it is these problems which refer most directly to the translation process in its essence, A model which does not take account of the information-content of language cannot provide a sufficient answer to the problems of a line of research whose first allegiance is to information. Of the questions we asked when discussing the "map" analogy, one is answered in part. The linguist's map will not do as it stands. We need other information which it does not contain.

Numbers of attempts have been and are being made to achieve Mechanical Translation using the methods and criteria of Structural Linguistics. Generally, they embody elaborate and sophisticated procedures for discovering the "formal" structure of the input text, and for ensuring that the output consists of "formally" viable sequences in the output language. However, the semantic procedure usually consists of a few more or less arbitrary rules for choosing one of a small number of output alternatives which are listed against each input word. I believe that neither this, nor anything like it will, in any sense, "do". Nearly all the intellectual criticisms of Bar-Hillel[1] can be leveled

1. Op. cit.

with justice and effect against such a program. It is the basic thesis of this paper that a program based on the methods and findings of Linguistics, that is, a syntactic program, will not do; that a program which tries to operate purely on a lexical basis, that is, a semantic program, will not do; and further, that these two working in parallel, that is, a syntactic together with a semantic program, will not do. The only hope for High-Quality Mechanical-Translation is a program which provides an area in which information provided by each of these methods is combined in the same form. I believe it is possible to show, and there is certainly no theoretic argument to refute this, that the information provided by each can be combined in an interesting and important manner with that provided by the other. Those who do not believe that the information-content of language can be represented by a "formal" system are many[1]. They abandon Mechanical Translation for the same reasons which caused "Structural" linguists to abandon semantics.

We have seen that translation does not depend on any "formal" correspondence, in our sense, between original and translation. Furthermore, we have seen that an approach which insists on regarding the semantic and "formal" aspects of language as different in essence, is unlikely to be the most productive for Mechanical Translation. Such a view would imply that the dichotomy between the semantic and "formal" aspects itself constituted a "formal" distinction, and this is manifestly not the case. The distinction does not therefore lie within the competence of the linguist to make, so that while he may agree not to take account of semantic criteria, he cannot exclude them from his system. He is unable to isolate, on "formal" grounds, the features of language, which have semantic content from those which do not. It can therefore be shown that remarks of the following kind are not valid, "The semantic content of an utterance is a property of the aggregate of the morphemes which make it up, considered

1. Y. Bar-Hillel, op. cit., in particular Appendix IV.

as an unstructured set." No linguist denies that semantic patterns carry a semantic burden of their own, however, it is not his business to recognise what this burden is. This is part of the interpretation and not part of the system itself. Consider a simple example. The two sentences

"Have you a car?"

and

"You have a car".

are composed of the same elements, but would correspond to different formulae in a linguist's calculus. Also, they have manifestly different meanings. In this case, the difference in meaning corresponds to a difference of syntactic pattern. In another language, it might correspond to a lexical difference, as in Latin, or to a simple intonational difference as in Russian or colloquial French. If it be claimed that this difference is trivial, and that the distinction I am making between "formal" and "lexical" is unclear, I account my point well made. This sort of phenomenon is a phenomenon well-known to all human translators. They know that an essential part of the translation process is a redistribution of semantic information between the lexical and 'formal' systems of the target language. Clearly, therefore, the successful Mechanical-Translation program must embody a stage in which such a redistribution can take place in accordance with the demands of the target language. This stage will be capable of representation by a calculus which can be interpreted either in lexical or 'formal' terms. Such a calculus I shall call an "Interlingua".

All the work on Mechanical Translation carried on at the Cambridge Language Research Unit has been directed towards translation using an Interlingua, and various types of Interlingua have been used in experiments performed there.

The Unit believes that the type of Interlingua which is likely to be most useful is the "Thesaurus"[1]. Examples of how such an Interlingua might be used have appeared

---

1. See M.M. Masterman, "What is a Thesaurus?"

from time to time in the Unit's publications[1]. However, since it is my avowed purpose to demonstrate the necessity of an Interlingual stage where grammatical and lexical information are reduced to a common form, weaker than either, so that they may be subsequently redistributed, let us briefly consider an example of how a Thesaurus might be used for this purpose. Consider the sentence

"I finished reading the book before dinner"

and the Russian translation

Я прочиал книгу до обеда

Other translations might, of course, have been possible, but this may very well be the one required in a particular context. What we have to decide is how the same result might have been obtained mechanically. The point of interest, of course, is the translation of the English word finished. The Immediate-Constituent method can be made to produce a perfectly unequivocal "formal" analysis of the sentences, enabling them to be bracketted in this manner:

((I (finished reading))(the book))(before dinner)

and

(Я (прочитал книгу))(до обеда)

The formal structure is thus similar except in that two of the Russian words are translations of English bracket-groups. One case is simply accounted for by the fact that Russian has nothing to correspond to the English articles, and the information they convey is usually lost in the translation of an individual sentence. The other case is less straightforward. "To read" has been rendered by its normal equivalent читать modified grammatically to include the ideas expressed in the English word, "finished". Clearly any number of words

---

1. In particular, Appendices to R.M. Needham & E. W. Bastin, "A New Research Technique for Analysing Language"; A.F. Parker-Rhodes and C. Wordley, "Mechanical Translation by the Thesaurus Method using Existing Machinery", Journal of the Society of Motion Picture and Television Engineers, Vol. 68. (1959).

might have been used in the place of "finished" and rendered in the same way. How can a transference of this kind be achieved mechanically. Let us rewrite the Russian sentence as follows

(я ((читал perfective) книгу)(до обеда)

The output sentence can then be regarded as a transformation of this. A correspondence between the English "finished" and the Russian perfective aspect is thus made explicit. Grammatical features such as aspects, moods, cases etc. are entered in the Thesaurus together with items more usually thought of as lexical. If we compare the entries for the English "Finish" and the Russian perfective, we shall see how they could each be made to contribute to the selection of the other in a translation program[1].

	English "Finish"	Russian Perfective
44 Disjunction	1	1
50 Whole	0	1
66 Beginning	0	1
67 End	1	1
70 Discontinuity	1	1
106 Time	1	1
111 Transience	0	1
113 Instantaneity	1	1
119 Different time	0	1
134 Occasion	0	1
140 Change	1	1
142 Cessation	1	1
292 Arrive	1	1
729 Completion	1	1
731 Success	0	1
732 Failure	0	1

The entries for both in the rest of the thousand sections are 0.

---

1. The numbers and words on the left-hand side of this table represent sections in Roget's Thesaurus of English words and phrases. A "1" in either of the other columns represents an entry in an M.T. dictionary under that section. The classification used by Roget is by no means ideally suited to the present purpose, but it is at present the nearest approximation in English.

Not only have these entries *a* great deal in common, but one is totally included in the other. Had this not been the case, it would have been necessary to find a word to express the residue. As it is, we are bound to choose either the imperfective or the perfective for the verb, and since the word which is grammatically most closely associated with the verb is included in one of these, it can be left out altogether.

I have given this example as a rough indication of how such a procedure might operate rather than as a recommended general method. It is only if some procedure of this kind is used, that we can hope to avoid translating, "Mr. Britling sees it through" by "M. Britling y voit clair" - and that was done by a human.

The explicit exclusion of information-content from among the features Linguistics recognises in Language has at least one other important consequence. Freed from the need to find a place for semantics in his description, the linguist is less constrained in the choice of the models he may create. Very few linguists would claim that there is one correct analysis of any text or one set of criteria for such an analysis(1). Y. Bar-Hillel(2) has tried to show that the Phrase-Structure, or Immediate Constituent Model is "inadequate". Now, one may not claim that a given model is or is not adequate unless one has first said for what it is or is not adequate(3). This, in my opinion, Bar-Hillel has not done. He does not relate his remarks in any but the most general way to Mechanical Translation, and it is, from what he says, by no means clear what it would be like for a model to be adequate for this purpose.

1. "But grammatical analysis is still, to a surprising extent, an art; the best and cleverest descriptions of languages are achieved not by investigators who follow some rigid set of rules, but by those who, through some accident of life-history, have developed a flair for it." C.F. Hocket, "A Course in Modern Linguistics" (p. 147)

   "Some linguists believe that grammatical analysis has become completely objective, but this is not true", ibid.

   "In this country, the tendency has been to eschew the assembly-line process, and to leave the application of the theory to the genius of the individual craftsman." W.S. Allen, "On the Linguistic Study of Languages", (p. 15).

Footnotes to Page 13. continued.

2. "Decision Procedures for Structure in Natural Languages", originally a talk given before the "Colloque de Logique", Louvain, September 1958, and printed in a revised version as Appendix III to op. cit.
3. "Whether a grammatical description of a language is satisfactory or not depends in part on the use we want to make of it". C. F. Hocket, "Two Models of Grammatical Description", Word, 10 (1954).

Before one begins to speak of the adequacy or inadequacy of any technique, even within the apparently restricted frame of reference of Mechanical Translation, one has to have answered a number of questions.

First, are we to consider the adequacy of this (or any other) technique to a strictly bilingual method of translation or is an interlingual method envisaged? If we are to consider an interlingual method, do *we* propose using this technique in the input or output parts of the program?

Let us, for the minute, discuss the relevance of decision procedures to Interlingual Mechanical Translation, One of the arguments most frequently adduced in favour of the interlingual method is economy. Translation from any one of n languages into any other, which with bilingual methods would require $\underline{n(n-1)}$ procedures, can be carried out with interlingual methods with 2n procedures [1]. The fact that these expressions are not $\frac{n(n-1)}{2}$ and $\underline{n}$ respectively shows that an essential difference is recognised between an input and an output procedure. Now it is by no means clear that the type of linguistic analysis appropriate to one will be equally appropriate to the other. An input syntactic procedure is required to recognise in any given text, just so much information as is required in the interlingual stage, that is, just so much information as will effectively distinguish it from any other meaningful sequence of the same lexical units. For this purpose, it cannot be shown that only one possible method of analysis will serve or even that the method chosen should yield results which correspond in every case to an intuitive analysis. On the grounds of economy, it is desirable that as many

---

1. Bar-Hillel (op. cit) asserts the fallaciousness of this argument, for if one of the natural languages were used as an interlingua, the expression would be $\underline{2(n-1)}$. This, however, would not have the principle advantage I have claimed for an Interlingua. I shall also adduce other arguments.

of the processes involved in the translation should be included in the interlingual stage of the program. It is unlikely, for example, that any criterion so strong as a test for "sentencehood" need be applied in a previous stage, and it is not even certain that, in every case, the analysis need be carried so far up the hierarchy imposed by the Immediate Constituent Model as to include all the intuitively recognisable sentences in the text. An empirical criterion of "sentencehood" would be, as I think Bar-Hillel has convincingly shown, almost impossible to find. Indeed, decision procedures of this kind have no place in Mechanical Translation, and probably not in language study at all [1]. The point which has not been made enough, and which is central to this discussion, is that syntactic structure in natural languages is not a well-defined notion. Any attempt, therefore, to show that such structures cannot be found seems to me ill-conceived and of small interest. Is this not what W.D. Preston means in his often quoted and variously interpreted remark:

> "Structure is a series of statements. The structure of a given language does not exist until it is stated." [2]

So far, no mention has been made of Transformation grammar which has recently been made part of Linguistics by Zellig Harris and Noam Chomsky.[3] There can be no doubt that this adds greatly to the power of Linguistics and that it will contribute to the ultimate success of Mechanical Translation. However, once again, we have to consider the place it is likely to take in the program,

---

1. The place of decision procedures in Mechanical Translation is treated more fully by K. Sparck Jones (ML87).
2. IJAL (1948), Cf. also:
   "...linguistics assumes no categories in rebus, no system inherent in the material and awaiting discovery." W.S. Allen, "On the Linguistic Study of Languages" (p.14), Professor Allen's lecture is particularly interesting in this context.
3. See in particular Noam Chomsky, "Syntactic Structures".

Broadly speaking, the set of transformations derived from a single kernel are patterns in which the same basic lexical units nay be combined, but with differing meanings or, at least, different stress. Seen in this way, the method seems to answer to many of the requirements of interlingual Mechanical Translation, for it stands squarely across the dividing line between grammar and lexicon. The process of constructing output text can be envisaged as one of selecting words and transformations from an inventory where they were listed together against some form of interlingual equivalents. We are told that given a language "L"

> "The grammar of L will...be a device that generates all the grammatical sequences of L and none of the ungrammatical ones."(1)

- and if the grammar is one of the Transformational type, there seems to be every reason for hoping that it may be caused to generate grammatical sequences in the output stage of a mechanical program. But will it be useful in the input stage?

We have said that the syntactic part of an input procedure is required to recognise, in any given text, the information required in the interlingual stage. In this, we made no reference to "grammatical" or "ungrammatical" sequences. However, if we are to glean from the syntactical patterns such semantic information as would enable them to be differentiated in the Interlingua, something analogous to the sort of output procedure suggested will have to be included. We require to be able to look at grammatical structures as a whole, not just at their constituent parts and to name them so that there is some method of looking them up in an inventory. Now, it is claimed that the Transformational method of analysis is able to resolve questions of ambiguity which are beyond the power of other techniques, and to this extent, at least, it is superior. Utterances which, though differing in meaning, would be analysed similarly or ambiguously by, say, the Immediate-Constituent method,

---

1. Chomsky, op. cit. (p. 13).

can be shown to be derivatives of different kernels. Thus if S is made to represent a certain sentence, $A(k_1)$ a transformation of the kernel k_1 and $B(k_2)$ a transformation of the kernel k_2, we may find in our list of permitted sequences;

$$S \leftrightarrow A(k_1), \text{ and}$$
$$S \leftrightarrow B(k_2)$$

so that S, which, by other methods would be ambiguously analysed, are here shown to be derivatives of different kernels. But seen from the point of view of an input procedure, this constitutes a statement of the ambiguity and not a solution of it. In the output stage, it is reasonable to suppose that $A(k_1)$ or $B(k_2)$ is "given", and the problem is only to find S. There is no ambiguity here, but in the input stage, S is the "given". There is one reason why Transformation grammar is more appropriate to the final than to the early stages of a translation program. Here is another. It is extremely difficult to apply these methods directly to a "raw" text, especially by strictly mechanical means. The length of a sequence listed as a kernel, or a transformation, is arbitrary and unlimited, so that their recognition entails the same difficulties as are encountered in the recognition of morphemes (or "chunks") in a sequence of words[1]. Nevertheless, if the recognition process is combined with an analysis of a different kind, by Immediate-Constituent methods, for example, many of these troubles can be made to appear less formidable. But if the analysis is to be performed by other methods in any case, are we still sure we require to recognise the transformations? Experiments with the "Thesaurus" approach to Mechanical Translation give reason to hope that much of the information which would be derived from a Transformational analysis could be obtained in the Interlingual stage. This would be clearly preferable for the reasons already stated.

---

1. See R. H. Richens and M. A. K. Halliday, "Word Decomposition for Machine Translation".

I am convinced, therefore, that the Transformational model of language will prove an invaluable tool for its original purpose, that is, to "generate" grammatical sequences, but that some other method, resting on a weaker set of assumptions about the input text, will be required for the first analysis[1]. It is reasonable to expect of a mechanically produced translation that it should consist of only viable sequences of the output language, but I see no reason to set quite so close a restriction on the texts which can be accepted as input. It is an essential part of the "interlingual" hypothesis that where the syntactic routine fails to give a unique analysis of a given piece of text, the missing information will be supplied by the semantic routine, and vice versa.

The arguments in this paper have been based almost entirely on the assumption that High-Quality Mechanical Translation will be achieved, if at all, with the use of an Interlingua, and that this Interlingua will not be a natural language, but an artificially constructed system. T*O* justify this, the essential nature of the translation process and an argument based on the requirements of economy have been invoked. There are two other arguments: one practical and the other theoretical.

First, the practical argument, which has been put by R. H. Richens[1]. Assuming the principles to have been established upon which Mechanical Translation can be made to operate, the formidable task of preparing dictionaries

---

1. I have recently been giving time to devising a punched-card routine for applying Immediate-Constituent methods, in varying forms, to texts. The procedure is extremely simple, and I see no reason why they should not be made to yield as much information as would be required for Interlingual Translation. I hope to publish the results of this work in the near future,

2. "Interlingual Mechanical Translation", The Computer Journal, Vol.1 (1958).

and inventories for all the languages to which it is proposed to apply it still lies ahead. To prepare these requires people with, not only a knowledge of the principles, but also of a source and a target language. Where these languages are English, French, German, Russian and the like, the problem is not insuperable. But shall we find it so easy to have these dictionaries and inventories prepared for such pairs of languages as Estonian and Hindustani, or Welsh and Georgian? Moreover, will it not be between just such pairs of languages that Mechanical Translation will do its greatest service? Surely it is when the rice growers of a remote Indian village, urgently need advice from Japanese experts to save their rice crops, that Mechanical Translation will really repay in concrete terms the work that has been done on it. If the method is interlingual, we need not find people knowing both these languages to prepare the necessary material, but only one of each. This is possible.

The second argument is theoretical, and is concerned with the use of an artificial, as opposed to a natural language, as an Interlingua. Since languages have grown up in diverse ways as products of different civilisations, a perfect translation is not possible. Not only is information inevitably lost in translation, but irrelevant information is gained; noise is introduced. The words of the target language carry with them associations which are necessarily different from those of the source language while the original associations are, to some extent, at least, lost. However, our Interlingua need not reflect any civilisation in particular; we hope it will be a weak enough calculus to carry ideas and associations from widely different languages and civilisations. If the Interlingua is an artificial language, we have at least some control over this. We may reasonably expect, therefore, that the amount of noise introduced into the system will not be of the order of twice the average amount introduced in normal translation, but slightly over the average amount. In fact, the result may be expected to be superior to a translation of a translation by human translators.

It would be gratifying to finish a paper such as this with a laconic phrase summing up the relationship between Linguistics and Mechanical Translation. Too much of both of these subjects lies still in the future, and I am not a prophet. That Linguistics has a relevance to Mechanical Translation does not require to be demonstrated, but there is, I believe, a danger in thinking that their relationship is straightforward. The danger lies here; that Linguistics says at once too little and too much about its subject matter, to be applicable, as it stands, to Mechanical Translation. Structural Linguistics has limited itself to certain well-defined aspects of language-study, and it has succeeded in describing these in terms of a strong set of postulates. These are too strong for our purposes. We do not require all the information which can be squeezed out of a language and presented in terms of a formal system. We are not principally interested in what features the text before us has in common with other texts in the same language. We require, first and foremost, to discover what is <u>different</u> in this text; what is there in it which makes it <u>this</u> and not some other text? What, in fact, is it trying to say? These are questions which linguists do not try to answer; the ideal of Mechanical Translation will only be achieved when they are answered fully and mathematically.

To sum up, the failure of linguistics to supply all the equipment necessary for high-quality Mechanical Translation can be accounted for by consideration of the nature of Linguistics. It thus becomes apparent even before the possibility of interlingual translation is envisaged. But to supplement the inadequacy of this equipment requires an Interlingua. Once you have constructed this device, which on the argument of this paper, is one which you have to have anyway, the supplementary arguments for using it for Interlingual Translation, with differing linguistic aids at the input and output stages, become strong.

Martin Kay
Cambridge Language
Research Unit.