MECHANICAL TRANSLATION

DEVOTED TO THE TRANSLATION OF LANGUAGES WITH THE A I D OF MACHINES

News 1

Cambridge Language Research Group Issue
Abstracts, discussions and three papers in full 2

Graphic Linguistics and its Terminology
R. A. Crossland 8

An Electronic Computer Program
for translating Chinese into English
A. F. Parker-Rhodes 14

Preprogramming for Mechanical Translation
R. H. Richens 20

Bibliography 29

Published at the MASSACHUSETTS I N S T I T U T E OF TECHNOLOGY

MECHANICAL TRANSLATION

DEVOTED TO THE TRANSLATION OF *LANGUAGES* *WITH* THE AID OF MACHINES

VOLUME THREE, NUMBER ONE JULY, NINETEEN FIFTY SIX

News

INTERNATIONAL CONFERENCE

An international conference on mechanical translation will be held at M.I.T., October 20, 1956. Papers will be presented from a number of groups working in the field. Those interested in attending should write to Victor H. Yngve, Room 20-B-101B, M.I.T., Cambridge, Mass.
Further details will be sent when they are available.

The conference is being sponsored jointly by the Department of Modern Languages and the Research Laboratory of Electronics, M.I.T., with the support of the National Science Foundation.

During the four days preceding the conference, there will be an opportunity for workers in MT to meet and discuss informally more technical papers which each will present as a basis for discussion. Those interested in contributing papers to the conference or to the discussion should write immediately.

GEORGETOWN GRANT

Prof. L.E. Dostert of the Georgetown University Institute of Languages and Linguistics has received a substantial grant for linguistic research related to mechanical translation. The grant, from the National Science Foundation, is for a period of one year starting in the fall. The research will be primarily, but not exclusively, focused on Slavic and English. Prof. Dostert will be assisted in this work by a group of structural linguists including Drs. Mueller, Pantzer, Brown, Garvin, and Austin; a group of eight graduate assistants in linguistics, and several translator-lexicographers.

CONFERENCE AT NAMUR

An International Congress of Cybernetics was held June 26 to 29, at Namur, Belgium, under the auspices of the Government of the Province of Namur and UNESCO. The second of four sections, "Semantic Machines," was under the chairmanship of Louis Couffignal, director of the Laboratoire de calcul mécanique de l'Institut Blaise Pascal (Paris). The topics discussed at this section included mathematical machines, machines for translating (theory of language, programming, special machines), universal logical machines and learning machines.

Cambridge Language Research Group

Meeting at King's College, Cambridge, England, August 2-4, 1955

THE EDITORS have turned this issue of MT over to the Cambridge Language Research Group for publication of the proceedings of their meeting at King's College, Cambridge, August 2nd to 4th, 1955. The meeting was attended by the speakers listed below and ten others interested in the field.

The Cambridge Language Research Group was founded in 1954 by Margaret Masterman. It is a compact, informal group of research workers from diverse faculties whose common interest is the scientific study of language and the application of the results of this Study to mechanical translation, a field of research still outside the official curriculum of any English university. Margaret Masterman, the chairman of the group, has specialized in the logical analysis of language. Among the other members of the group are: R.H. Richens, one of the first investigators to devise actual machine translation procedures and a contributor to Machine Translation of Language; A.F. Parker-Rhodes, who is studying the application of algorithmic procedures and lattice theory to the syntactical problems involved in machine translation; Lady Hoskyns and E.W. Bastin, mathematicians, and M.A.K. Halliday and R.A. Crossland, specialists in descriptive linguistics. R.H. Thouless is the president of the group.

The present research program of the group includes further development of linguistic analysis, the elaboration of a general program by means of which a passage in any one natural language can be translated into any other language via an algebraic mechanical interlingua, and exemplification of the latter by a pilot Italian-English translation project.

LIST OF SPEAKERS

E.W. Bastin*, Theoretical Physicist, Fellow of King's College, Cambridge.

R.B. Braithwaite, Knightbridge Professor of Moral Philosophy, Cambridge.

J. Bronowski, Director of the Central Research Establishment, National Coal Board.

J. Chadwick, Lecturer in Classics, Cambridge.

R.A. Crossland*, Lecturer in Ancient History, Newcastle.

J.R. Firth, Professor of General Linguistics, London.

M.A.K. Halliday*, Assistant Lecturer in Modern Chinese, Cambridge.

* Member of the Cambridge Language Research Group.

C.W. Kilminster, Lecturer in Mathematics, King's College, London.

M.M. Masterman (Mrs. Braithwaite)*, Chairman of the Cambridge Language Research Group, Lecturer, and Director of Studies in Moral Sciences, Fitzwilliam House, Cambridge.

A.F. Parker-Rhodes*, Statistician, Cambridge.

R.H. Richens*, Assistant Director, Commonwealth Bureau of Plant Breeding and Genetics, Cambridge.

T.J. Smiley*, Mathematical Logician, Fellow of Clare College, Cambridge.

C.L. Stevenson, Professor of Philosophy, Michigan.

R.H. Thouless*, President of the Cambridge Language Research Group, Reader in Educational Psychology, Cambridge.

LINGUISTIC ANALYSIS AND TRANSLATION

J. R. Firth

Abstract

SIMULTANEOUS oral translation works best when cognate languages are concerned. There tends to be considerable mutual assimilation between such languages when science, religion or even politics are involved. An agenda also assists.

The case of a Russian addressing the recent Orientalists' Conference in English was mentioned. The intonation, gestures, etc. of the speaker conveyed information that the text of his speech did not.

It could be argued that complete translation is theoretically impossible.

An instance in which a scientific abstractor admitted that an adequate abstract was easier to make than a full translation was quoted. Abstracting could be regarded as one of the modes of translation.

The problem of determining the efficiency of translation is clarified if different modes and objectives of translation are admitted.

The American structuralists claim, though hardly with justification, to exclude meaning from their analysis of language.

Hjelmslev distinguishes between content and expression but has not produced any detailed linguistic description on this basis. The concept of purport as an entity common to different languages is obscure. Hjelmslev emphasizes the dependence of content on expression. The value of Hjelmslev's distinction between reference, designation and signification in descriptive linguistics is doubted. Commutation is applied by Hjelmslev to changes in expression paralleled by change sin content.

A distinction was drawn between the language under description, the language of description and the language of translation. The London School avoids translation meanings whenever possible.

Z. Harris was criticized for his use of translation meanings.

Attention is drawn to Malinowski's ethnographic analysis. Malinowski distinguished between interlineal, running and free translation. He sought out explanations of the use of words from speakers of the language concerned, and collected series of mutually exclusive words coming into the same semantic field.

The concept of primary meaning or core of meaning was criticized, using "ass" as an example.

The study of restricted languages was recommended. Examples are Malinowski's language of garden magic, the language interposed in mathematical texts and liturgical language.

DISCUSSION

Chairman: R. H. Thouless

PROF. BRAITHWAITE pointed out that the different kinds of restricted language in French and German mathematical books derived from the different conventions of writing mathematical treatises that had developed in the two countries.

PROF. FIRTH quoted Aviation Japanese as another example of a restricted language.

He also mentioned the question raised by ancient Indian grammarians as to whether the meaning of a word was implicit in the first syllable or only when the entire word had been spoken.

The value of studying restricted languages, even such restricted samples as long words of classical derivation in English, was reemphasized.

The difficulties of translation from exotic languages were described.

PROF. BRAITHWAITE presumed that "exotic" implied "in respect of a particular language or languages".

PROF. FIRTH said that English was exotic to Melanesians. No question of primitiveness was involved.

MISS MASTERMAN asked whether one could state that a language A was exotic to a language B in respect of a purpose P.

MR. RICHENS thought that "exotic" was a relative term and could be applied to any pair of

dialects or languages but in different degrees.
DR. PARKER-RHODES thought that one could translate exotic languages by appropriate use of loan words.
PROF. FIRTH thought that closely related languages such as English and Dutch would be most suitable for pilot schemes in machine translation.

He referred to the phonological level of meaning in poetry which could hardly be translated from, say, English to French. Syntax and stylistics presented different modes, again, to the translator.
MISS MASTERMAN inquired whether the number of levels of translation was to be regarded as determinate or indeterminate.
MR. RICHENS disliked the term "level" as applied to translation if this implied serial order in efficiency. Word-for-word translation conveyed some information lost in a smooth translation.
PROF. FIRTH disliked the mixture of different modes of translation in word-for-word translation.
DR. HALLIDAY said that the categories used in linguistic description depended on the aim of the description. It was possible to conceive of an interlingua with a universal scheme of categories. At the grammatical level, he suggested that the descriptive categories could best be regarded as points in an n-dimensional manifold.

PROF. FIRTH explained the distinction between structures consisting of interrelated elements such as word classes and systems of paradigmatic units giving values by commutation. He thought that special categories might be required for translation.

MR. RICHENS expressed concern at the lack of consideration given to the ideas expressed by words. He referred to structurally ambiguous passages in Japanese botanical writing which could only be translated accurately if the scientific ideas concerned were known to the translator.
MR. CROSSLAND suggested that a sufficient amount of context would suffice instead.
PROF. FIRTH said that the Principle of Mutual Expectancy would operate in these as in other passages.

MISS MASTERMAN suggested that the conflict between the mathematical approach, conceived in terms of "naked ideas", and the approach of descriptive linguistics was only superficial and that the two approaches were in fact complementary.
PROF. FIRTH agreed with Mr. Richens that translation via descriptive linguistics would be intricate.

He elaborated the application of the notions of structure and system, pointing out that these prescinded from any concept of time sequence.

Dr. Bronowski paid a tribute to the late Prof. Haloun whose interest in the logical structure of ancient Chinese had been one of the sources of inspiration of the Cambridge Group.

NEW TECHNIQUES FOR ANALYZING SENTENCE PATTERNS

M. Masterman

Abstract

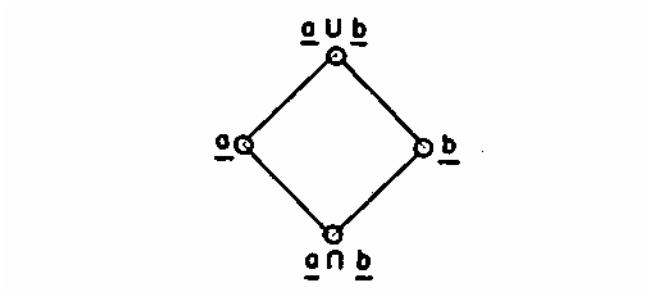
THE TWO NEW techniques that appear promising for sentence analysis are based on combinatory logic and lattice theory, respectively. It is thought that the work of the Cambridge Group, in which these two techniques are being used, gives greater promise for the establishment of a satisfactory theory of language than Hjelmslev's approach.

Hjelmslev is believed to have been misguided in attempting to construct a general deductive theory from analytic descriptions of particular texts. Also Hjelmslev's philosophy of science is markedly positivist in tone and ignores other views of science such as the predictive. Moreover, Hjelmslev's method of treatment has led him to develop a proliferating terminology in which the distinctions between the terms tend to be logically vacuous; he seems to be in danger of repeating Whitehead's massive mistake of avoiding the need for formal theoretic deduction by retreat into a private language.

The logicians of the Cambridge Group, influenced at an earlier stage by Prof. Haloun, regard Chinese as a favorable language for machine translation experiments since it is logically less variegated than many others, and is

built up from a unit the tzu[4] or concept more fundamental than the "word". Special attention has been given to a particular sentence selected at random from a Taiwan botanical journal.[1] This has been analyzed in terms of a single binary operation, in accordance with the principles of combinatory logic, on the assumption that the concatenation (a(b)), leading up to the focus of emphasis, constitutes the primary method of combination of language symbols. This operation can be interpreted "b is qualified by a", or, more generally, "b is limited by a". The grammatical distinction between full tzu[4] and form tzu[4] is believed to be assimilatable to the logical distinction between arguments and operators, the latter functioning, in this new application, as indicators of the type of bracket required, and thus as combinators, while the former are the elements. Thus, in the experimental sentence, the form word tzu[4] 'however' sets up brackets between the sentence (B) and some preceding sentence (A) to give the arrangement (A(B)). Some degree of parallelism between A and B must be presumed.

However, the multicombinatorial bracketing of long sentences is apt to become highly intricate, and it seems that chain configurations, lattices or networks that can be embedded in a lattice may be simpler to handle. Thus, qualification could be expressed as a 2-element chain lattice and the double relationship a and/or b and both a and b like by the Boolean lattice



Sentence operators can therefore be regarded as corresponding to the ∪ symbol in the case of such conjunctive tzu[4] as chi[2] (and) or to the ∩ symbol in the case of such product-forming tzu[4] as chih[1]. The experimental sentence is on p. 28.

DISCUSSION

Chairman: J. Bronowski

MR. RICHENS pointed out that two types of configurations were being used by the Group, one in which the elements were the actual words occurring in the sentence analyzed, and a second in which the elements were purely semantic, prescinding altogether from the lexical and structural peculiarities, such as active or passive constructions.

PROF. FIRTH said that linguists had no objection to methods of ordering or reordering the words of a sentence, nor to the distinction between argument and operator. The notion of dependence of words on each other, however, was of questionable validity especially when primary, secondary or tertiary dependencies were distinguished.

Many of Hjelmslev's linguistic views derived from those of French linguists and of sociologists of the school of Durkheim.

MISS MASTERMAN amplified, from the logical point of view, her criticisms of the large corpus of new terms introduced by Hjelmslev.

PROF. FIRTH reviewed some of Hjelmslev's terms and expressed uneasiness at the abstruse nature of his "schemata".

MISS MASTERMAN would not allow that Hjelmslev's system was deductive though she mentioned that Prof. Braithwaite was prepared to argue that it could be regarded as the limiting case of a deductive system.

The CHAIRMAN thought that Mr. Richens' view that active and passive voices could be interchanged in translation was rather radical.

He went on to consider the different significances of "mass" in "I throw this mass" and

1. The Cambridge Language Research Unit has issued a further Progress Report which is dated January 7, 1956, and which constitutes a supplement to the Conference Report of August, 1955. This supplement consists of a paper by Margaret Masterman setting out comparatively four analyses of the Taiwan botanical sentence here referred to, the analyses given of the sentence being i) in terms of a specially constructed combinatory notation ii) as a lattice iii) by applying to the sentence a combinatory logical system (cf. on this, R. Feys, "La Technique de la Logique Combinatoire", Revue Philosophique de Louvain, 44, pp. 74-103 and 237-270, 1946.) iv) by applying to the sentence a skeletal computer program which uses analyses i) and ii). This skeletal program has been designed by A.F. Parker-Rhodes and M.T. Hoskyns from a pre-program designed by M. Masterman. (Note by M. Masterman)

"I lift this mass". This, he said, was a basic kind of problem with which a theory of language would have to deal.
MR. RICHENS said that he distinguished between transfer of meaning and transfer of structure in translation.
PROF. STEVENSON wondered what happened to the "standard forms" of traditional logic if such grammatical forms as active and passive voices were interchanged.
PROF. BRAITHWAITE said that logical "standard forms" were useful only for a particular pur pose and were not to be regarded as in any sense absolute.
PROF. STEVENSON observed that for many purposes "crude translation" was all that was required.
DR. PARKER-RHODES pointed out that any change in a linguistic text entailed a change in the information conveyed. The practical issue was the amount of content to be transferred.
PROF. STEVENSON asked whether it was not easier to transfer structure than to alter it.
MR. RICHENS said that it would be easier in some cases.
PROF. FIRTH said that what had been called transfer of structure was merely the result of accidental parallels in the structures of the two languages concerned.

He could envisage an interlingua built up as a bridge between two languages each of which had been adequately surveyed from a descriptive point of view.
MR. RICHENS did not wish to regard the bridge merely as a construct from the two languages concerned. He thought that an interlingua could contain independent semantic information and pointed out how, in scientific translation, accuracy depended on knowledge of the science concerned.
PROF. FIRTH said that a specification of extended collocations would answer instead of special scientific knowledge.
MR. BASTIN inquired to what extent West Indian Pidgin could be regarded as a bridge between Chinese and English.
MR. CROSSLAND discussed the correspondence of the grammatical systems of voice, noting, in particular, the diverse translations of the Greek middle voice.
PROF. FIRTH gave examples to show how parallelism of structure would be expected to increase as increasingly cognate languages were compared.

GENERAL MATHEMATICAL PROBLEMS INVOLVED IN MECHANICAL TRANSLATION

E. W. Bastin

Abstract

THE GENERAL problem is how to reduce the number of possible meanings of a sentence composed of words each of which has an extended range of meaning. A possible solution is to regard words as indeterminate units which are increasingly determined as the context is built up. The rules to be applied to the indeterminate units can be regarded as a "language theory" which should be expressible in algebraic form. The language may make use of lattices in which the ordering relation corresponds to the succession of words, the lattice points to the possible words of the discourse and the junctions to the logical connections. An actual sentence is then represented by one particular path through the lattice. There will also have to be a starting and finishing point.

The conventional parts of speech are represented by characteristic lattice subconfigurations.

At any particular stage in the enunciation or writing out of a sentence the set of possible completions constitutes a spread which becomes more determinate as information accumulates. The notion of spread is an extension of Brouwer's concept.

The set of possible translations of a sentence also constitutes a spread which it is desired to reduce. It might prove possible to introduce a number of subroutines, each representing a language theory, into a computer and devise an interaction mechanism whereby the language theory best able to reduce the spread of possible translations could be ascertained by successive approximation.

DISCUSSION

Chairman: J. Bronowski

DR. THOULESS said that there appeared to be some inconsistency with regard to the lattice bonds, which seemed to denote both logical relations and the succession of words in the sentence.

The CHAIRMAN asked whether the sentence "the dog bit the cat" could be put into lattice form by way of example.

MISS MASTERMAN pointed out that this could only be done when some specification had been made of the nature of the restricted language out of which this sentence had been taken.

PROF. FIRTH said that indeterminacy could be reduced by considering sufficiently restricted languages.

PROF. STEVENSON asked why the progressive determination of a sentence should be envisaged as proceeding from the beginning to the end. Given the end of a sentence, it was possible to restrict the range of indeterminacy of the beginning.

MR. RICHENS said that given any words of a sentence, not necessarily contiguous, the range of indeterminacy of the rest was restricted. There is no need to regard progressive determination as proceeding linearly in either direction.

Graphic Linguistics and its Terminology

R. A. Crossland

DURING the past thirty years great advances have been made towards making the study of language a science, but leading linguists have been mainly concerned with spoken language. There has been a certain tendency to suggest that the study of written documents should always be subsidiary to that of some spoken idiom, or even that it is bound to be less scientific than that of spoken idioms, and perhaps not a proper part of "linguistics" at all.[1]

These suggestions should be opposed. "Linguistics" should include the study of written languages as well as that of spoken; the former study can and should be as scientific as the latter, and it needs its own terminology which should be basically independent of that of the study of spoken languages. Much confusion, and some mistrust, if not antagonism, among linguists would seem to have resulted from lack of agreed distinct terminologies for the two studies, which might well be called respectively phonic and graphic or epigraphic linguistics.[2]

The problems of graphic linguistics are probably best approached through consideration of what writing is. A script may be defined as a system of visual symbols whose purpose is to convey the thought of one individual or group to another. Writing is often treated as a means of representing a spoken utterance or utterances by visual symbols, but this is not its primary purpose, except where phonetic or phonemic transcription in linguistic work is concerned. Representation of actual, contemplated or imagined utterance is a particular mechanism for conveying meaning by graphic signals, one whose convenience lies in the small number of signs required. The adoption of a particular form of it, alphabetic writing, in Western Europe, has led to its being widely regarded as the normal and natural mechanism, and some of those who have discussed the analysis of systems of writing have tended to write as if they were all more or less satisfactory systems of phonemic transcription of utterances. This attitude leads to or supports the view that the study of written documents should always be subsidiary to the study of some spoken idiom, or as an extreme to the idea that "texts" are not "language".[3] One must leave to psychologists the question whether it is possible to read or write without some thought of phonic[4] realization, whether based on a known spoken idiom or not. But it can hardly be denied that the users of a system of graphic communication may develop for it conventions of vocabulary and grammar which differ from those of any spoken language which they use, or on which the system was originally based. A group of texts showing similar conventions of grammar and vocabulary may reasonably be termed a "written language".[5]

Most of this will probably be accepted by the majority of those concerned with the study of spoken languages, though in some cases with the proviso that the study of written language should be considered a discipline separate from "linguistics" and "philology." Such differentiation, however, has the disadvantage of tending to dissociate the study of the spoken form of a Ian -

---

1. Cf. W.S. Allen, "Phonetics and Comparative Linguistics", Archivum Linguisticum 3, (Glasgow), 126-36.

2. Choosing between graphic and epigraphic here involves a problem common when technical terminology is devised, whether to use the term which is etymologically the most natural, in spite of its currency in non-technical language in another sense. For epigraphic, cf. A.F.L. Beeston, Transactions of the Philological Society, 1951, 1-26, where it means 'of the inscriptions'.

3. Cf. Allen, op.cit., pp.132, 136.

4. As phonetic is now generally used of description of utterances or segments of utterances according to the manner of their articulation, a more general term to cover all studies concerned with spoken language is required, and phonic seems suitable. The use of phonics proposed by J.R. Firth, Trans, of the Phil. Soc., 1951, 84, has not become widespread.

5. Or a "written dialect", if its relation to another group with closely similar conventions is under consideration.

guage from that of the written, where both forms exist, a development particularly undesirable in the case of semantic studies. "Linguistics" should include the analysis and study of the mechanism of both spoken and written languages, while "philology" should be used of studies of the content of written texts, in particular for historical or literary ends. This usage is in fact normal in American English, and corresponds to German use of Sprachwissenschaft and Philologie. "Philology" and "graphic linguistics" will overlap to some extent, especially in semantic studies, but there is a clear distinction between the two in purpose.

Graphic linguistic study, as well as phonic, may reasonably be called "descriptive" or "structural" if its procedures are appropriate. An analysis of the conventions of a class of texts may be termed "descriptive" if it is not shaped by a preconceived notion of what they should be; "structural," if it aims at determining signifi - cant oppositions.

Recent work in phonic linguistics has esta - blished a terminology for phonetic and phonemic description of spoken languages, and recently suggestions have been made for a similar terminology to be used in analysis of written languages.[6] None has yet become generally accepted, how - ever, and those proposed seem unsatisfactory in so far as they are based mainly on the partly phonemic, alphabetic scripts[7] of Western Europe and are not easily applicable to scripts of other types. The analyses which they imply are in some cases not purely graphic, as they reflect the function of the written signs or the conventions of their combination in representing phonic features of spoken languages.

The terminology now most used in Britain in describing spoken languages permits description at three levels: phonetic description of a single utterance, phonetic description of a number of utterances, and phonemic description, which may be defined for present purposes as description on the basis of contrasts significant to normal users of the language in question. Distinction is made, for example, between a sound which seems to require definition as "the audible result of a single emission or intake of breath or closure or opening of speech organs by a particular speaker on a particular occasion"; a sound-class - any group of sounds, as just defined, which an investigator associates, perhaps provisionally, in analyzing the phonetic structure of a language, for example, on grounds of phonic similarity or occurrence in similar contexts; and a phoneme, which for convenience may be defined as a sound-class differentiated functionally from others.[8]

It has been recognized that graphic linguistics needs a set of terms similar to sound, sound-class and phoneme in the technical language of phonic linguistics. It would seem to need at least a term for a sign, modification of a sign or feature of arrangement in a particular segment of a particular document; one for a group of similar signs, modifications or features classed together, provisionally or permanently, in graphic analysis; and one for any such group which appears to contrast significantly with another or with zero. Graph or sign suggests itself for the first, graph-class or sign-class for the second, and grapheme for the third. To illustrate the use of these proposed terms, a in a particular written word; for example, class , in

---

6. See D. Abercrombie, "What is a 'letter'?", Lingua 2, 54-68; P. Diderichsen, "Nye bidrag til en analyse af det danske skriftsprogs struktur", Selskab for Nordisk Filologi, Arsberetning for 1951-52, (Copenhagen), 6-22; E. Pulgram, "Phoneme and Grapheme: a parallel". Word 7, 15-20; H.J. Uldall, "Speech and Writing", Acta Linguistica 4,11-6; J. Vachek, "Some remarks on writing and phonetic transcription", Acta Ling. 5, 86-93. Diderichsen's article seems particularly important.

7. Cf. Pulgram, Word 7. 15; " . . . . each alphabet has a certain number of . . . . classes of symbols . . . ." (my underlining).

8. In passing, the choice of sound as a term for the first concept in the publications of most members of the London University School of Oriental and African Studies seems unfortunate. The creation of new terms in technical language is preferable to use of current ones with new artificially restricted meanings. Moreover, sound has long been used in philological and linguistic literature with an accepted sense: the range of "sounds" (in the restricted sense just mentioned) which normal speakers of a language known only from written documents are thought to have produced in pronouncing - "giving phonetic realization to" - a word-segment represented by a given phonic grapheme (cf. "the sound f in Lat. filius", the meaning of which is clear enough). However, a term for the restricted concept to whose expression some would limit sound in the technical language of linguistics is certainly needed. Perhaps phone would serve; cf. Pulgram, Word 7,15.

this present text, would be described as a graph; all small a's of similar formation in a document or group of documents as a graph-class. Only full examination of how a script is employed in documents under consideration--analysis of its structure, that is to say--will indicate which graph-classes should be termed graphemes. For example, graphic analysis of a sufficient number of documents in modern English would lead to three varieties of written A being distinguished as graph-classes; a, a, A. Structural analysis would probably require the first two being considered to form, together, a single grapheme. since, except in special texts, such as phonetic transcriptions, they never contrast significantly in the same document. Capital A would probably have to be considered a grapheme in written English. Its occurrence at the beginning of sentences may be considered not to involve significant contrast with small a, since sentence division is indicated by the full stop. But there are cases where the use of capital or small a initially is the only graphic indication whether a person, place or group of persons or places is referred to, or some more extensive concept: cf. the Archers and the archers.

A principal difficulty of graphic analysis will be to decide whether certain features should be considered independent graphs or graphemes (according to the level of analysis) or not. In the case of most scripts there will be an obvious division into what may be called provisionally unitary graphemes and graphemes of arrangement or modification. The simplest case is offered by a linear phonemic script, which uses gaps to indicate word-division. In this case each letter will be a unitary grapheme representing a segment of a spoken or imagined word. Sequence of unit graphemes from right to left in scripts using the Latin alphabet, will be an arrangement grapheme representing temporal order of enunciation of the segments which they represent. Juxtaposition of unitary graphemes, at less than certain intervals in normal texts, will be an arrangement grapheme indicating that the segments represented constitute a word. Italicizing to indicate emphasis is an example of a modification grapheme. Description of graphemes according to their function in scripts which are only partly phonic in principle will be a good deal more complicated. It might be fairly simple in a fundamentally ideographic script--Chinese is the only example, I think, apart from the earliest Sumerian.[9]

9. The Chinese script is the obvious example. Others are the earliest Sumerian and Egyptian, and the Mayan.

The differentiation of unitary graphemes and graphemes of arrangement or modification should be a fairly simple process. It will often be more difficult to decide whether a particular symbol is to be regarded as an independent grapheme or not. Decisions will have to be made on grounds of ease of recognition, or with regard to the ideas of those who normally use the script in question. For example, it is arbitrary and a matter of convenience whether we analyze the Sanskrit signs usually transcribed -ra, re, -r, (final position only), ri, ru, pa, pe. -p (final only), pi, pu, as eight separate graphemes, or as six, k and p, modified by a grapheme zero (indicating following a), and graphemes representing following i, following u and absence of following vowel. If, in analysis of a linear script, superlinear or sublinear symbols are treated as graphemes, it will presumably be necessary to differentiate them from unitary graphemes and graphemes of arrangement or modification.

Differentiation of graphemes on the basis of the manner of their employment in the script to which they belong is the only proper differentiation in a descriptive study of a written language. Differentiation of graphemes according to the manner in which they are used to represent concepts and their nexus will be necessary when the history of a script or the interaction of written and spoken forms of a language is studied. One may then want to make a distinction, for example, between phonic graphemes, which indicate a concept by indicating more or less accurately its oral realization in a spoken language, and what are generally termed ideograms, but which for the sake of symmetry within the terminology one might better call idea graphemes, concept graphemes or notional graphemes.[10]

A complex terminology would be needed to describe e.g. Babylonian cuneiform, which is partly syllabic, partly ideographic.[11]

From the point of view of mechanical translation, the following seem important:

10. Logogram should only be used of a sign representing a particular word. It would be incorrect, for example, to apply it to the Sumerian sign No. 172 in P.A. Deimel, Sumerisches Lexikon, which represents in different contexts bil, "burn", and izi, "fire". A purely logographic script would be impracticable for most inflected languages. The number of signs required would be prohibitive.

1. Written texts can be scientifically described and analyzed without reference to any spoken form of the language in which they are written or to the spoken language which the script in which they are written originally was devised to represent.
2. Problems of ambiguity resulting from homography in written texts are not likely to be more frequent or more serious than those which result from homophony in a spoken language.
3. No system in regular use will represent the nuances conveyed by emphasis or intonation in a spoken language, but this is not a serious objection to mechanical translation of written documents of the type in use in most modern civilized countries. In the written forms of many languages, nuances, of the type mentioned, in the spoken forms are conveyed by alternative means, and an individual may quite well express his ideas in the written form of a language, (or even in a dialect or foreign language which he does not speak) more precisely than in the spoken idiom which he normally uses.
4. Although a phonemic text may be regarded as an abstraction of utterances, it is probably better to regard written and spoken forms of a language as different realizations of concepts and their nexus than to regard either as on a higher level of abstraction than the other.

ritten texts ca

DISCUSSION

Chairman: R. H. Thouless

PROF. FIRTH reviewed American work on the descriptive analysis of written texts. In Arabic script the different forms of initial, medial and final letters afford a criterion of the limits of a word lacking in Roman script. It was pointed out that even when speech was being analyzed, this was always reduced to a text. The proper use of the terms "phonic", "phonetic", and "phonological" was explained.

DR. PARKER-RHODES wondered whether it would be possible to stop epigraphic analysis at the level of the word, each of which could then be regarded as a unit.

PROF. BRAITHWAITE said that the occurrence of unspaced texts indicated that too much emphasis should not be placed on the word as a unit.

MR. RICHENS said that analysis below the level of the word was essential in machine translation. It is not feasible to construct dictionaries in which each inflected variant is treated as a single word, therefore it is necessary to break up inflected words in mechanical translation.

With regard to Sanskrit script, a special epigraphic problem is involved in respect of the characters written in reverse order.

MISS MASTERMAN wondered whether one could distinguish between graphemes representing single sounds, syllables, and words.

PROF. FIRTH pointed out that some African languages could be written both conjunctively and disjunctively.

PROF. BRAITHWAITE made a comparison between Mr. Crossland's treatment of signs and the logical distinction between "token" and "type".

MR. HALLIDAY discussed in what sense it was correct to speak of Chinese as a syllabic language.

MR. CROSSLAND said that the syllable was largely a reflection of the method in which the speakers of a language analyzed their own speech.

MR. CHADWICK said that the syllabic script of Mycenaean Greek was probably devised for some other unrelated language.

MR. RICHENS referred to the cumbersome renderings of occidental languages in Japanese kana. He wondered what significance should be attached in epigraphic analysis to modes of expression found in the written but not in the spoken language.

---

11. A syllabic grapheme may be defined as one representing a phonic segment which those who devised a syllabic or partly syllabic script thought they could distinguish when they attempted to analyze words of the language which they spoke, for graphic representation.

12

THE THEORY AND PHILOSOPHY OF LANGUAGE

J. Bronowski

Abstract

THE FACT that translation is possible is an important philosophical datum, the significance of which has not yet been properly appreciated. Denial of the possibility of translation is best regarded in the same light as solipsism.

Two philosophical schools have concerned themselves especially with translation within a language, the school represented by Carnap and that of the Oxford analysts. The views of both these schools were critically reviewed. It was thought that Carnap had overestimated the possibilities of complete axiomatization in mathematics. This topic was discussed with reference to a theorem relating to algebraic surfaces; these are three proofs of the theorem, but the axiomatic systems underlying these have little in common.

It appears to be often the case that a theorem can be deduced from an unproved hypothesis (e.g., that of Riemann) and can also be demonstrated independently. Yet hopes of utilizing the independent theorem to establish the un proved hypothesis are usually disappointed. This is characteristic of a branch of mathematics which is still growing.

It is believed that no empirical system which is still making discoveries and a fortiori no living language can be fully axiomatized. Perhaps a dead language could be.

The role given to personal exegesis by the Oxford analysts was critically examined. While it is true that analysis reveals difficulties in the translation of sentences involving the speaker, these objections do not apply with the same force to impersonal statements.

The analogy between the output of machine translation and poetry, especially Chinese poetry, was pointed out. In each case the reader supplies connections not made explicit in the text.

The view that words have a zone of uncertainty that is narrowed by contextual relations was endorsed. On the other hand, there is frequently redundant overdetermination as well.

Language is a process of quantification of individual words in their different contexts. The picture of language as a lattice fixed by interlocking relationships is an attractive one.

Seen from the philosophical angle, it is only difficult translation that is really worth doing.

DISCUSSION

Chairman: R. H. Thouless

MISS MASTERMAN agreed that special difficulties attended translation of sentences involving the speaker. The Oxford analysts have shown that the notion of a common proposition underlying statements in different languages can not be accepted without considerable qualification.

PROF. FIRTH observed that the Oxford analysts had derived some of their concepts from descriptive linguistics.

MR. SMILEY pointed out that the sort of languages discussed by logicians are very far removed from natural language.

PROF. BRAITHWAITE thought that Dr. Bronowski had underrated the extent to which mathematics could be axiomatized.

MR. SMILEY thought that language presented a practically virgin field for applied mathematics.

MISS MASTERMAN said that it was incorrect to assume that mathematics necessarily held the key to the situation. In the history of science, there is a tendency to underrate the importance of the presystematization stage.

MR. RICHENS did not see why the difficulties in axiomatizing pure mathematics should necessarily apply to language.

DR. BRONOWSKI replied that since language mentioned the empirical data of science, the difficulties of axiomatizing the latter would apply also to the former.

PROF. BRAITHWAITE pointed out that axioms could be replaced by rules.

MISS MASTERMAN said that operative principles can usually be demonstrated even though axiomatization had not been accomplished.

MR. RICHENS thought that axiomatization of language might be feasible at the word-class level. He pointed out that the range of human emotive expression could be regarded as comparatively static over the last few thousand years and might therefore prove susceptible of axiomatization.

DR. BRONOWSKI thought that the completeness with which a science could be axiomatized was in inverse proportion to its vitality.

MR. RICHENS said that, in respect of language, axiomatization might prove possible in some fields if not over the whole domain. He agreed with Dr. Bronowski that axiomatization was not likely to prove the most effective approach to the problem of translation but maintained that the impossibility of axiomatizing language had not been demonstrated.

DR. BRONOWSKI said that if Mr.Richens agreed on the practical issue, he was content. He believed that his inference, although only presented heuristically, was valid.

DR. PARKER-RHODES said that if a general theory of language were to emerge, he would expect that translation would provide the starting point.

An Electronic Computer Program for Translating Chinese into English

A. F. Parker-Rhodes

General Considerations

The procedure known as translation consists in the expression, through the medium of the target language, of that information which is conveyed by the text in the source language. We shall not consider here the conveyance of anything apart from "information" in the narrow sense.

We have further to consider that the information latent in the source text may not all be relevant for the purposes of the exercise. Languages differ considerably in the kinds of information which they consider as "relevant." For example, in English we cannot convey any verbal concept without at the same time adding information about when the action took place relative both to the moment of speaking and the moment of reference. In Chinese on the other hand all this extra information is regarded as irrelevant. Differences between relevant and irrelevant information are not only due to differences in linguistic habit, but may be due to the common human tendency to include irrelevant matter rather than to risk leaving out anything of importance. Theoretically, a "sufficient" translation could be defined as one which conveyed all the relevant and none of the irrelevant information. But this would be a poor aim for a computer program, (a) because when the same "irrelevancies" are present in both languages, trouble is saved by letting them pass, and (b) the rigorous pruning of, for example, English tenses, would lead to an undesirable "pidgin" effect which can in fact fairly easily be avoided.

We therefore aim instead at carrying over all the details which do not add to the operational labor involved, and as little as is necessary to inform the target text with a minimum of elegance.

Catataxis

The required information is supplied in the source text in the form of a simply-ordered series of symbols. In the case of Chinese, these symbols are "characters." I shall say nothing here as to how these characters are to be "recognized", except to emphasize that from social and moral considerations the process ought ultimately to be mechanized, and not relegated, as some have suggested, to a semi-skilled operator, which would merely replace a highly educated translator by a less developed type of worker.

The symbols in the source text, together with their ordering-relations, contain all the information available. The semantic content of these two kinds of item may be interchanged as between source and target languages. For example, we have:

Chinese	$ting^1fang^2tsu$	$fang^2tsu\ ting^1$
English	top house	top of house

the relation which is expressed in the Chinese text by an ordering relation, is expressed in English by the addition or omission of a word. In the case of closely-related languages such cases may be relatively few, but in general the effect of this interchangeability will be to make the distinction between "words" and "word-orderings" a nuisance. One stage of our process must therefore be to reduce all items of information, however conveyed in the source, to a common form. This stage I call "catataxy".

There are two main ways of doing this. The first is the "lexical", the second the "algorithmic". Lexical methods aim to list all the relevant forms, be they words or word-orderings, and to record for each listed item an appropriate equivalent in the target language. [An example of the application of lexical methods to catataxy is described by Mr. Richens]. On the other hand, algorithmic methods seek to prescribe rules, analogous to the rules which we learn in the elementary processes of arithmetic, whereby the significant word-orderings can be discovered and represented by numerical symbols (like those by which we convey, in the computer, the "meanings" of the separate words); and subsequently introduce further rules, to convert these symbols into others which will indicate the word order required by the target language. The method of catataxis which I have worked out is of the algorithmic type.

Metalexis

Before I describe these methods in further detail, it is necessary to consider in some detail what form those symbols will take, by which the source text is represented in the machine. These symbols will be obtained as the output of a dictionary, whose input is provided by the signs delivered to it by the reading device. Here at once we come upon what is probably the most difficult question in machine translation. How are we to sort out, from the great variety of "meanings" capable of being attached to a given word, the one appropriate to the given context? The difficulty is only partly allayed by the fact that we shall be using, in practice, restricted languages. Even in the most restricted form of Chinese, for example, chung[1] will have, among its possible meanings, "middle," "during," and "China," while fang[4] for example will require 5 or 6 "basic" equivalents.

Two considerations can be applied to choosing the appropriate meaning in such cases: contextual and grammatical. The use of contextual criteria really amounts to further restriction of our restricted language as we go along. It will consist in practice of arranging to store in the computer a series of indications of context, drawn if possible from individual words; for example, a word such as "thrilling" could be counted as excluding the context "technical papers", while a word such as "influorescence" would carry much weight in excluding, for example, "navigation". In connection with this system, each of the alternative meanings contained in a dictionary entry will carry a "key", arranged to "fit" (in a sense defined according to the elementary operating of the machine) the "lock" in which the accumulated contextual information is stored.

As regards the grammatical criterion of choice, each alternative might carry an indication of the kinds of other words it can be associated with. For example, chung[1] after a noun preceded by such verbs as tsai[4] or tao[4], and/or followed by ti(chih), may safely be rendered by "among" or (with time-words) "during". These words can themselves be identified by special signs -- "word-class indicators*. The procedure here, therefore, will involve entering at first for each word a provisional word-class indicator, indicating the W.C.I.'s of all the alternatives not excluded by the context criterion, and then, as subsequent words are read in, the provisional W.C.I.'s must be read through to see what possibilities they exclude in regard to the grammatical contexts. It may well be necessary to go through the whole sentence twice before the full range of information is brought to bear on each word.

At the end of this process, if rightly programmed, we shall have selected a single alternative for each word of the source text, and this alternative will be represented by (a) a code sign, which the output dictionary will turn into a word of the target language, and (b) a W.C.I, being another code sign conveying the grammatical functions possible to this word in the source language in the given context. These W.C.I.'s will provide the raw material for catataxis.

The Kind of Algorithms used in Catataxis

The program by which catataxis is carried out must begin with a master-routine which will identify the various W.C.I.'s, and direct the computer to turn to the further algorithms appropriate to each case. The identification of W.C.I.'s is done by subtraction: they are arranged in. the numerical order of their respective symbols and suitable quantities subtracted in turn from them; the computer will then recognize each by how soon the resulting number becomes negative. The processes applied to each word-class vary considerably. In each case, the objective is to build up, from the original W.C.I., a symbol which indicates not only the word-class of the word, according to an appropriate grammatical analysis of the language, but also its relations, so far as they are relevant, to the other words in this particular sentence. This symbol I have called a "taxon"; it is worthwhile to consider in some detail what form these taxa will take.

In principle, this is largely arbitrary; different methods may well *be* found convenient for different purposes. We have heard already of two possible methods of organizing sentences in mathematical terms, and the program I have proposed makes use of both "brackets" and "lattices" (or rather, chains). The only problem, in using a procedure of this type for the construction of taxa, is to select a suitable method of representing the chosen mathematical forms by the binary numerals which alone the computer can handle.

The binary representation of brackets is based in my system on the assignation of a particular binary place to each pair of brackets. Thus, in the accompanying example, in the taxa A, the square brackets[ ] enclosing the verbal group have in common, for all the enclosed words, the digits 10 in the 1st two places. The round

Table

showing the proposed arrangement of entries in the Input Dictionary

The linear order is that to be realized on the input-feed of the computer, and need not be reproduced on (say) dictionary cards.

Location (of not more than 20 dgts)	Nature of the Contents	
0	Code sign (as from reading device)	
1	Provisional W.C.I. (= sum of all others)	
2)	Sign for "expect compound" or	
)		
3)	"check for compound"	
4	Key for grammatical lock	1st alternative meaning
5	Space (either gramm. or context)	
6)		
)	Key for context lock	
7)		
8	W.C.I. for given alternative) or sign for	
9	Meaning for given alternative) cpds	
10 - 15	The same for next alternative	
4 + 6n)	Control combination to stop reading-in until	
)	the metalexis (sorting of alternatives) is	
5 + 6n)	finished.	

brackets, enclosing the "complex group" (Halliday) qualifying the verb tsou^3, have in common the additional 3 digits 001; the small brackets containing the compound hua^1 yuan^2 have a further 11, which they share with their postpositive noun li^3 (in practice, such a compound as this would be separately entered in the dictionary). In this system A (which is not the one finally adopted) one can further perceive that the relation between verb and postverbal noun is indicated by the change of 01 into 11 not only at the level of the main sentence (in the 1st two binary places), but also in the subsidiary group (in the 5th and 6th places). This, in practice, is a quite unnecessary refinement; it is possible to work out the structure of all sentences completely without this information, and to abandon it makes possible much shorter taxa and simpler programming.

I therefore turned from the system exhibited in A to that of B. Here only the smaller brackets are retained, the larger brackets being replaced by a pattern of "chains". These are represented by prefixes, in which words belonging to one chain have a 1 in a prescribed position. In the example, the main-sentence chain is represented by a 1 in the second place of the prefix, and the complex-group chain by a 1 in the first place. The word tsou^3 at which the two chains join has a 1 in both places, thus showing the structure of the sentence just as clearly and much more economically than by the bracket-notation.

Having decided on the representational principles to be used in our taxa, we have to devise the necessary algorithms to derive the required binary forms from the given series of W.C.I's. This involves, first, an appropriate method of predetermining the W.C.I.'s, and, second, a set of routines for distinguishing the various groups of words which require to be recognized in the taxa. It will be noticed that in our examples the W.C.I.'s themselves form generally the last part of the finished taxon, the earlier digits being added by the algorithms. [The words yuan^2 and li^3 are exceptions, since their endings 1 and 101 receive an extra 1 to show that yuan is the second element of a compound] .

To show the sort of form our algorithms take, this last is an appropriate example.
First, when we find any taxon assuming a form identical with its predecessor, then the required algorithm is called in. Thus, at an appropriate stage, we arrange for the taxon to be subtracted from its predecessor; if the result is 0, the

		Taxa A	B	W.C.I.
t'a^1	he	010000.100001	01.01.1	01.100001
[(tsai^4	at	100010.01	10.10.01	10.01
(hua^1 }	garden	100011.1	10.11.1	01.1
yuan^2) }		100011.11	10.11.11	01.1
li^3)	in	100011.1101	10.11.1101	01.101
tsou^3]	walk	101000.1	11.10.1	10.1
lu^4	road	110000.1	01.11.1	01.1

N.B. The points are entered for ease of reading only; in the computer each digit has its fixed place and such aids are not needed.

taxon stands and is entered in the place of its W.C.I.; but if the result is 3420, we have to arrange (i) to find the last 1 in the next taxon (or the last 101 if the W.C.I. has this ending), (ii) to add a 1 in the next binary place. The taxon thus amended must be substituted for its W.C.I. In most cases, we have to add the new digits at the beginning, and to facilitate this the digits forming the W.C.I. are placed in such a position that they do not have to be shifted at all during the formation of the taxon. Often, however, a taxon has to be altered in the light of subsequent words of the sentence.

Anataxis

When all the operations required in Catataxis have been completed, all the W.C.I.'s supplied in the original input have been replaced by taxa. Each taxon is thus followed, in the storage locations of the machine, by a code sign representing its chosen "meaning" in the target language. Thus every significant feature of the given sentence, whether a word or a word-ordering, is now represented by a binary numeral. This series of signs has now to be so manipulated as to indicate correctly the order of words required in the target language.

It might in some cases be possible so to arrange the system of taxa so that they should give, by their own numerical order, the order of words ultimately required. However, this would necessitate the use of a different system of catataxis for each target language as well as for each source language, and also the algorithms required would be more complex than they need be. Thus, it is convenient to use a separate set of algorithms to alter the taxa, so as to achieve the required re-ordering.

This set of algorithms I call Anataxis, since it puts together again that which catataxis takes to pieces. (If the procedure is based on lexical methods, no separate stage is required for anataxis). As regards programming, it is simpler and shorter than Catataxis, and presents no special problems, at least as between Chinese and English which have rather similar word-orders; the main points are that in English the qualifying phrases, of the kind which in Chinese end in ti^4 or chih^1, are placed after the word qualified instead of before, and that adverbs can always (though if style is to be sought, should only sometimes) follow their verbs.

In the example given above, the group in the outer round brackets needs to be placed at the end of the sentence, and this would be achieved in my program by (i) spotting it as a qualifying group (by the sequence of prefixes 01,10,11,01, separating 10,11 as the required group) and (ii) altering these prefixes so as to read, in this case, 01,11,10 (the 11 covering both the 10 and 11 of the original sequence). In other cases, other parts of the taxa must be altered; e.g.:

man^4	10.001	slowly	10.101
man^4	10.0011		1011
		becomes	
tsou^3	10.1	walking	10.0
cho^1	10.101		10.001

which, on arranging in numerical order, gives "walking slowly". The necessary change consists in interchanging 0 and 1 in the third place (of those here represented) from the left.

Anaptosis

When the target language is inflected (unless the inflections have fairly exact correlates in the source language) a further stage is required after Anataxis, in which the required inflections are added to the otherwise incomplete word-forms. With Chinese as the source language no assistance at all is provided in this direction, as this language is entirely uninflected. With English as the target, the difficulty is increased by the related (but logically distinct) circumstance that the required inflections mostly express logical categories which Chinese usually ignores, such as number and tense.

In my programming essays hitherto I have been content with rather crude solutions to the problems of anaptosis. Thus, I have suggested inserting "the" before all nouns where the Chinese gives no indication to the contrary (such as is afforded for example by ko[4], chih[1], etc.). Likewise, I have expected that an appropriate "blanket" tense would be acceptable in most "restricted" contexts; for example, in scientific papers, all facts may be put in the past simple, and all opinions and hypotheses in the present. The insertion of plurals can be based on the presence of particular key words. As regards case, the only distinction which appears in written English is the genitive -s, which I propose to replace everywhere by "of".

These elementary expedients would hardly serve for a more highly inflected target language, and for these anaptosis would probably have to be combined with anataxis in a single but relatively complex program.

Output

What is left in the storage of the computer when the stages of catataxy, anataxy, and anaptosis have been completed is a sequence of "words" in the order left by the anataxis routine, each of which consists of a taxon and a "meaning". The latter will have been modified so as to include sufficient information to determine the inflectional forms required, (though in a highly-inflected target language the space needed for this may be too much to be accommodated in the same location as the main "meaning" code-sign).

The taxa, however, have now served their purpose and may be cleared or overwritten, so that their places could be occupied by the additional indications required,

The last stage of the process of translation may now begin: it consists in reading-out the contents of the still relevant locations, in their present order (which is that of the target language), to a suitable output dictionary which will convert the coded "meanings" directly into alphabetic signs capable of actuating a teleprinter which will write out the target text sentence by sentence. This may be done by whatever output mechanism the given computer may be filled with. Perhaps punched teleprinter tape would be the most convenient medium.

The output dictionary need not contain any of the complications of that used for input. The latter is required to carry the necessary information for metalexis, and this process cannot be put off, since it is (in general) necessary for the determination of the W.C.I.'s which are themselves necessary for catataxis. At the output stage, however, all that is required is to decode the meaning, already determined by the code-sign which the input dictionary has supplied. Therefore, the output dictionary will work on a one-to-one basis and be correspondingly simple in design.

One of the main difficulties in mechanical translation is likely to be that of checking. In mathematical computations it is a regular and usually necessary practice to include sundry checks in the main programs. The nature of the translation process precludes this possibility. The best that can be done is to examine the output to see that it is not nonsense; this is hardly a sufficient check, but it is rather unlikely that an error in the computer would be such as to lead to "sense" other than the correct sense.

DISCUSSION

Chairman: Prof. Braithwaite

DR. HALLIDAY queried the distinction between anataxis and anaptosis.

DR. PARKER-RHODES said that it was mainly a matter of convenience and might not apply in other pairs of languages.

MR. RICHENS pointed out that two distinct types of information were required for inserting the definite article, one based on the structure of the base sentence and the other on the structural characteristics of English.

PROF. FIRTH agreed that no satisfactory English grammar existed for the purpose under discussion. He analyzed the spectrum of meaning in dictionary entries. The range of tense in languages is often narrower than commonly supposed.

MR. CROSSLAND discussed the problem which grammatical agreement would present, in connection with the proposed stages of anataxis and anaptosis.

Preprogramming for Mechanical Translation

R. H. Richens

TRANSLATION is a species of communication in which the set of symbols adopted by the communicator is changed into another set of symbols before reception. It is possible to argue that all communication involves such a substitution of symbols and that communication within a single language is merely a limiting case of translation. For present purposes, however, we shall confine the scope of discussion to translation between different spoken or written languages.

We have next to inquire as to what remains invariant in translation. If we try to convey the maximum significance of the symbols of the base language, it is clear that a great deal is involved: gross meaning, the subtler overtones, deliberately concealed meanings, manifestations of the subconscious mind, the sound of the base words or their appearance in script, metrical characteristics, etymology, the associations engendered by the communication, the statistical characteristics of the communication as a sample of the output of a particular author or period, and the pleasure or otherwise engendered by communication in an informed or cultivated recipient. It is obvious that a mere fraction of all this comes over in any translation and hence we derive the notion of translation as a scaled process. We translate at various levels and in respect of various characteristics. An additional limitation on the precision of translation is provided by the peculiarities of the target language which may contain no symbol for an idea in the base language, a frequent occurrence in the case of exotic plants or animals, or no method of rendering an idea without adding an inaccurate qualifier, as in Chinese-to-English translation where the neutrality of the Chinese noun with respect to number cannot be preserved.

The notion of level or mode of translation is important. Machine translation has earned a certain notoriety for its indulgence in very low-level translation and its fondness for what has come to be known as mechanical pidgin. For certain purposes, however, such as locating allusions, low-level translation may be all that is required. Confusion only occurs if the mode of translation is not made clear.

We are now in a position to discuss the notion of preprogram. Machine translation depends on collaboration between linguists, engineers and an obscure set of people interested in the bridge territory between the two, where problems of logic and semantics arise. It is not to be expected that a person whose primary interests are linguistic will appreciate the nicer details of electronic circuitry. It is therefore important to develop procedures that are comprehensible to linguists and engineers alike and can be used as the basis for developing detailed programs for any particular machine. Such general procedures are referred to here as pre programs. Till now, the devices principally used for experiments in machine translation have been punched-card machines and electronic computers. It is possible that the best machine for machine translation as regards both efficiency and expense has not yet been devised. It is important therefore to develop procedures that are not tied down to any particular machine but which can easily be applied to a particular machine when required.

A question that is of considerable interest is the optimum combination of man and machine. It has come to be generally recognized that machine translation with intensive human pre-and post-editing is hardly worthwhile since this method is largely concerned with remedying the defects of the machine. A far more satisfactory concept is that of companionship. An efficient translating machine that can operate whenever required, can continue when its human partner is fatigued, can instruct its partner without the wearisome labor of consulting dictionaries and grammars, and can retire quietly into the background when the human partner desires to exercise his powers unaided qualifies in considerable measure as a good companion.

After these preliminaries, we can proceed directly to concrete problems.

The following convention will be used. A term in single quotes is used to represent the word in the target language of which the quotation is a common meaning.

For purposes of machine translation it is convenient to distinguish between the following operations:

1. Transfer of meaning.
2. Transfer of ambiguity.
3. Transfer of structure.
4. Injection when, for example, number is attached to a neutral Chinese noun.
5. Restraint, preventing the machine from excessive semantic analysis.

The first stage in machine translation is character recognition. There are three possible methods:

1. Complete human recognition in which a reader deals with a familiar script.
2. Incomplete human recognition in which certain visual characteristics of an unknown script are picked out.
3. Photoelectric recognition, using standard fonts.

This stage is of very considerable importance as far as the economics of machine translation is concerned, but is irrelevant to the subsequent operations and is therefore excluded from the preprogram.

The outcome of recognition is the conversion of the symbols of the base text into a functional equivalent such as holes in punched cards or teleprinter tape. Having obtained a functionalized text, the next stage is matching against a mechanical word-dictionary. This operation has been discussed in some detail by R.H. Richens and A.D. Booth[1], and I shall only refer to essentials now. Each word of the base text must be matched against the entire mechanical dictionary, searching backwards. In some cases, a presorting of the base text into alphabetical order will expedite this operation. Then, as soon as a dictionary word is encountered which is wholly contained in the base word, the equivalent or equivalents in the target language must be entered. Should there be a residue, i.e., if a base word is inflected, the residue must then be matched against the mechanical word-dictionary in its turn. In the Chinese sentence studied by the Group, affixes do not come into the picture.

A point not sufficiently considered in the earlier paper concerns languages such as Latin with different conjugations and declensions or like Welsh with initial mutation. In this case, when transferring an affix, or in Welsh, the body of the word after cutting off the mutable initials, an indication of the conjugation must be extracted from the mechanical word-dictionary. Then, when matching the detached component, the conjugation indicator must be matched simultaneously.
Thus Welsh nhroed will be decomposed into

nh	(t declension)	— no meaning
roed	(t declension)	— 'foot'

The result of this operation is the sequence of equivalents dubbed mechanical pidgin.

Matching against the mechanical word-dictionary, however, cannot be confined to the matching of single words. In most languages, irreducible compounds occur such as "cool off" which in contrast to "im-possible" cannot be analyzed into semantic components. Such irreducible compounds must be entered as such in the mechanical dictionary. Then, when matching a word which may be part of an irreducible compound, it is necessary to extract both the meanings in isolation and the meaning in combination . A second matching is then necessary to ascertain whether the other component of the potential compound is present. If this is not, the compound can be erased. If the other member of the compound is present, it may be possible to accept the compound without further operation. In the Chinese sentence under consideration, the chances of encountering yung[2]-chieh[3] 'dissolve' in which the components retain their isolated meanings are relatively low.

It may be necessary, however, as in the case of German separable verbal prefixes, to defer a decision as to whether an irreducible compound is present until the syntax has been analyzed.

Whenever a compound is accepted, the meanings of the components in solution must be erased.

Thus, to obtain an output in mechanical pidgin, the mechanical dictionary must contain the words or parts of words of the base language, irreducible compounds, the equivalents in the target language, and indications of conjugation. In order to translate at a higher level, a more elaborate mechanical dictionary is required.

There are two types of information that we can utilize at our next level, syntactical and semantic. In the sentence "the dog bites the cat", subject and predicate are distinguished syntactically; in the sentence "this plant has yellow petals", semantic analysis indicates a botanical rather

1. Machine Translation of Languages. New York 1955, p. 24.

Sentence	WC	R	A	Pre	Post	WC	R	A	Pre	Post	WC	R	A	Pre	Post	WC
'however'	Advs	1				Advs	1				Advs	1				
'this'	Dem	2	2			NPp	2				NPp	2				
'two'	num	3				of										
'entity'	N^2	4	2													
'of'	of	1														
'appearance'	N^1	1		the		NPp	1									
'and'	and	2	1													
'dissolve'	V^2	4		of	-ing											
'degree'	N^1	3														
'somewhat'	Adv	1				Adj	1				Adj	3		are		
'not'	neg	2														
'alike'	Adjv	3	1													

WC -- word class
R -- rearrangement
A -- alternative
Pre -- preinsertion
Post -- postinsertion
Adjv -- adjective adverb
Advs -- adversative
NPp -- plural noun phrase

Table 1 : Schedule for syntax analysis

than engineering significance for "plant". Syntactic information will be dealt with first since it appears to present rather less complex problems than semantic information.

In order to analyze syntax, it is convenient to allocate words to word classes. In some cases these can be parts of speech or parts of speech delimited in various ways. Sometimes, in the Chinese chi[2] 'and', in which "reach" is an alternative meaning, the word class will be the sum of "and" and "verb". There is nothing against using different categories of word classes for different pairs of languages, though a general unified scheme has some obvious advantages. It is useful to allocate some of the most frequent multipurpose words to one-member classes of their own.

For utilizing syntactical information the mechanical dictionary must contain expressions for the word class of each entry; this will take the form of a number or series of numbers for each word. When translating at this level, the preliminary matching process now results in the output of a sequence of word class expressions corresponding to the sequence of words in the base text. There are now various possibilities. Dr. Parker-Rhodes would use the word classes to provide material for a computing schedule based on a moderately restricted set of instructions. I take this as analogous to learning a foreign language by means of a grammar. The method suggested here is more analogous to learning one's native tongue, in which correct usage is arrived at by imitation over a long period with no conscious realization of rules.

The mechanical dictionary in the present method must contain a supplementary dictionary of word-class sequences. The sequence of word classes for a single sentence is then treated as a single compound or inflected word. This is decomposed into its constituents in the same way as the individual words are decomposed into stem and affix, that is by matching the initial component first and then proceeding to the next and so on to the end. It is possible that, in the case of word-class sequences, the front may not be the best place to start, at least in some cases. This is a matter for further investigation.

The mechanical word-class sequence dictionary contains the following data under each entry:

1. Word-class sequence.
2. Rearrangement instructions.
3. Alternative instructions.
4. Pre- and post- insertion instructions.
5. Word-class equivalent.

The result of the matching procedure against the word-class sequence dictionary is to generate a series of instructions and a new word-class sequence. The latter then provides the basis fora new cycle of matching against the word-class sequence dictionary. The whole procedure is repeated until a word-class sequence is generated that is wholly contained in the mechanical dictionary. The operation is then concluded.

The accumulated instructions can then be read off, the rearrangements made, alternatives eliminated, and the necessary insertions made. In the Chinese sentence, three reductional cycles were involved. The procedure is illustrated in Table I. The output reads "however the appearance and degree of dissolv- ing of these two entities are somewhat un- alike".

The information utilized so far has been syntactical. The semantic information is more difficult to process and what follows is merely tentative.

A possible method is to attach semantic indicators to significant words and to collect the indicators as one proceeds through a passage, using the totals to decide between alternative renderings of doubtful words. Thus "petal", "stem" and "pineapple" could be accompanied by indicators for "botanical". This might help to limit "plant" to its botanical rather than its engineering sense. As Dr. Thouless has pointed out, some difficulty might be encountered with a "pineapple-slicing plant", but in this case "slicing" might carry an indicator pointing the other way. I am not in a position to say how useful this method could be. It has the advantage of collecting information as the text is traversed. However, it is obviously an extremely crude way of mobilizing semantic information and I should therefore like to consider next a more difficult but more fundamental approach.

I refer now to the construction of an interlingua in which all the structural peculiarities of the base language are removed and we are left with what I shall call a "semantic net" of "naked ideas". These bear some obvious resemblances to the linguistic configurations discussed already.

The elements represent things, qualities or relations. I associate adjectives (usually monadic relations) and verbs (dyadic or higher relations) in the Japanese way.

A bond points from a thing to its qualities or relations, or from a quality or relation to a further qualification.

"black cat" is

cat ⟶ black

"The cat is on the mat" or

"The mat is under the cat" is

cat $\xrightarrow{1}$ on $\xrightarrow{2}$ mat

In asymmetrical relations, the bonds are not interchangeable.

"The dog bites the cat" can be represented as

dog $\xrightarrow{1}$ part of $\xleftarrow{2}$ teeth $\xrightarrow{1}$ contact $\xleftarrow{2}$ cat

contact ↓ much

If a different category of bond is used for doubtful or uncertain connections, a method of precisely delimiting the field of ambiguity is available.

Constructions of the type dog ⟶ part of ⟶ teeth are not used since this would assume the possibility and desirability of weighting the terms of dyadic relations in terms of "superiority" or "inferiority".

When the Chinese sentence studied by the Group is represented as a semantic net, the figure obtained is of considerable complexity. What is more, various deficiencies in the information provided by the sentence become apparent; for instance, no mention is made of the solvent, without knowledge of which the significance of "solubility" is vacuous.

This raises the question of "restraint". A translator is frequently under the necessity of reproducing ambiguities or inconsistencies in the base language by corresponding ambiguities or inconsistencies in the target language. If a machine is to utilize semantic data, it must necessarily analyze the semantic relations of the passage fed into it. If this analysis is carried too far, the base passage is in danger of such severe mangling that a readable output in the target language will not be obtained. Thus in the example quoted, a machine that indulges in semantic analysis will demand information on the solvent; if however, it is restrained to conform to the frailties of human nature, it should be possible to stop analysis at the level of the concept "solubility" and present the smooth inadequate output that a human translator is expected to provide. It might prove possible to arrange for a machine to translate at various levels of restraint so that the ordinary person and the logician can each be satisfied.

The semantic net thus represents what is invariant during translation. It can, of course, be transformed into a unique linear sequence for dictionary purposes, rather in the way that the structural formulae of organic compounds can be given linear codes for purposes of cataloguing.

The problem of extracting semantic nets from base texts is difficult and no general mechanical procedure has yet been devised. One possibility is to regard the words of the base passage as pieces in a jigsaw puzzle. Each word has a number of semantic properties - differently shaped protuberances in the jigsaw analogy - which fit in with some words but not with others.

Thus the relation " $\xrightarrow{1}$ see $\xleftarrow{2}$ " can only attach on the left-hand side to a human being or animal. Syntax already restricts the number of possible combinations; semantics limits the possibilities still further.

If syntax and semantics do not lead to a unique interlocking, we have an ambiguous situation. Ambiguity can be represented in a semantic net by introducing a second category of bonds, and can presumably be transferred to the target passage if so required.

The syntactical procedure discussed earlier in this paper dealt with a specific pair of languages. It is more satisfactory theoretically to go through an interlingua that is capable of expressing the nuances of all the languages considered in a translation program and is more adequate for logical analysis than any existing language. Such an interlingua would have the practical advantage of connecting such languages as Welsh and Japanese, where the labor of compiling a specific translation program would not be worthwhile. It is well known that two-stage translation via an intermediary language is unsatisfactory; this is only so, however, when the intermediary language is a natural rather than a universal language.

The semantic nets described above have an obvious bearing on the question of a universal interlingua. If the elements (ideas) are replaced by letters with an ideographic significance only, we have in fact an ideographic algebraic script with obvious potentialities for machine translation work. The elaboration of a system of ideographs for handling discourse is one of the current research projects of the Cambridge Group.

In conclusion, I would like to return to the notion of translation as a scaled process in which a selection has to be made of the amount of information to be transferred. It is only a further

step to the notion of translation as a limiting case of abstracting. In ordinary academic life, especially in science, abstracts are required far more frequently than full translations. In the future, the increased rate of publication is likely to make the production of abstracts far more necessary. It therefore seems that any procedure of selective transfer of ideas is likely to be of considerable future interest. Semantic nets have an obvious relevance in this connection. This paper had, as its object, a brief description of some of the work being done by the Cambridge Language Research Group on machine translation. This work has now reached the stage where one is beginning to dabble seriously in schemes for machine abstracting.

CONCLUDING DISCUSSION

Chairman: Dr. Thouless

MISS MASTERMAN reopened discussion on the application of lattice theory to language.

There are two main problems: the search for objective criteria in setting up the lattice interrelations, and the discovery of methods for utilizing the information conveyed by these interrelations for translation.

The sentence "the dog bit the cat" was then projected into a configuration. (See Fig. 1). Possible elements of the discourse, not found in the actual sentence, were also plotted.[1] The actual sentence could then be represented as a path through the network of connecting bonds obtained.

In the method of analysis outlined, form words or operators occur either at the top or bottom of sublattices. It follows from this that operators will fall into two categories: abstractive operators, that is, meets, which render the words they govern less determinate, and concretive operators, that is, joins, which do the reverse. Any element, if it occurs at the top or bottom of a sublattice, can become an operator. Predicates are envisaged as abstractive operators.

The question then emerged as to how the configuration obtained above could be converted into a single lattice by the addition of suitable extra bonds and elements.

DR. KILMINSTER saw no reason why a lattice should be required, as opposed to various other partially ordered sets. A configuration of bonds of two types (potential relations between the elements and the path of the actual sentence) could be treated mathematically without attempting to reduce it to lattice form.

MR. RICHENS asked what the elements represented in the field of linguistics.

MISS MASTERMAN replied that each element was a possible concept in a restricted language. There are thus named elements (on the sentence path) and unnamed elements which could be connected to make other sentences in the same restricted language.

MR. RICHENS said that he could attach no linguistic significance to the extra bonds required to convert the configuration into a lattice.

DR. PARKER-RHODES inquired why the "the" of "cat" and the "the" of "dog" were plotted separately.

MR. RICHENS said that the discussion was in danger of drifting into the problem of universals.

MISS MASTERMAN said that what was being sought for was an exact technique for analyzing the relations between universals.

DR. HALLIDAY said that each "the" could be ear-marked by structural criteria.

MR. RICHENS said the projection of adjectives above their nouns and verbs below was accidental. He preferred the Japanese view that these two word classes were associated.

MISS MASTERMAN developed the view that the "all" element in a linguistic lattice, as well as representing the join of all the elements, represented the point of maximum determinacy, corresponding to a generalized "yes", while the "null" element represented the point of maximum indeterminacy, corresponding to overlapping of all the elements, i.e., to a generalized "no".

MR. SMILEY saw no need for converting linguistic configurations to lattices as opposed to other partially ordered sets - or other systems.

---

1. There was a wide divergence of opinion at the conference as to the best lattice representation of this sentence. Subsequent investigation has clarified, and to some extent modified, the views originally expressed by the exponents of this application of lattice theory. These later developments are being described in a forthcoming publication. (This note by R.H. Richens)

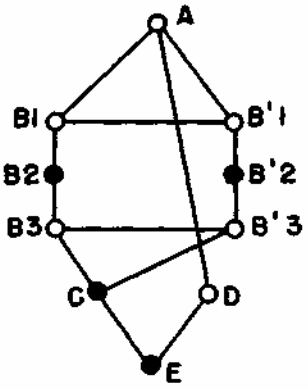


Fig. 1

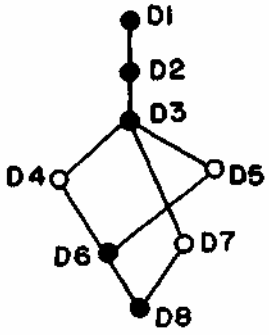


Fig. 2

Fig. 3

The Lattice Diagram for the sentence "the dog bit the cat" is shown in Fig. 1.
In this lattice the function of the various elements is as follows:

	Semantic Function	Verbal Representation
A.	Sign that the words included constitute one predication	Initial capital and final full stop
B.	Subject (or object) of sentence	The dog
B1.	Word defining a noun group	The
B2.	Word qualifying the group	absent in this case
B3.	Word able to stand for whole group	Dog
B'	Object (or subject) of sentence	The cat
C.	Situation involving "the dog" and "the cat"	absent
D.	Verb, connecting B and B'	Bit
E.	Word capable of standing for whole sentence	absent (but it might be represented by a initial "yes")

If it be desired to show further detail in the verb "bit", the element D could be itself expanded into a sub-lattice. This may be illustrated by the Latin equivalent "morsavit", as in Fig. 2. Here the elements are:

D1. Sign that the following segments all make up one verb
D2. Personal suffix, -it
D3. Sign that remaining segments all make up one verb stem
D4. Root of verb, mor(d)-
D5. Frequentative suffix, -sa-
D6. Represents completed frequentative stem
D7. Sign of perfect tense, -v-
D8. Represents completed verb.

In the case of a sentence with an intransitive verb, the structure is simpler, because there being only one noun the element C carries a meaning identical with that of B and so can be omitted as well. The simplified lattice for "the dog bites" is shown in Fig. 3. The two elements B and D could of course be expanded as above if required.

(This note by A.F. Parker-Rhodes)

He suggested an approach to the former based on finite geometry.

DR. BASTIN said that if a lattice were obtained, this had well-known correspondences with the simpler logical forms, so that a connection with Logic could be made.

MISS MASTERMAN recalled that a computer could be regarded as a distributed complemented lattice.

DR. HALLIDAY gave an example of a highly restricted language. He was uncertain as to how many occurrences of the same word should be regarded as separate elements. He favored distinguishing "dog" as subject and "dog" as predicate.

MR. CROSSLAND said that one should beware of obliterating the distinction between syntactic and semantic analysis. When a bilingual or multilingual dictionary is used, we step outside the field of descriptive linguistics altogether.

DR. HALLIDAY projected his sample of a highly restricted discourse into a single chain configuration; words were plotted twice if they occurred both as subject and predicate.

MR. RICHENS said that if the same adjective qualified two or more nouns, it must be plotted separately for each noun; otherwise it would be impossible, in some configurations, to reconstruct the original passage.

DR. PARKER-RHODES wished to know what possibilities there were of transforming the configuration representing the total set of relations of a base passage to the set of relations representing the configuration of its translation in a target language.

MR. RICHENS said that the mere possibility of translation implied that such a transformation was possible; however, he was of the opinion that semantic analysis was necessary in addition to structural analysis. He supposed that transformation without semantic analysis might be possible in some cases at the word class level.

DR. HALLIDAY said that the configurations illustrated two sets of conditions restricting the occurrence of lattice bonds between any given elements. The first set was grammatical and the second collocative.

The CHAIRMAN brought the discussion to a close and referred to the stimulating results that had been obtained by bringing linguists, mathematicians, and logicians together at the present meeting. A tribute was paid to Miss Masterman, the founder of the Cambridge Research Group, for the ultimate inspiration behind the meeting.

THE EXPERIMENTAL SENTENCE

The Chinese sentence used to exemplify the Group's work is as follows, typical meanings for each character in isolation being added in single quotes.

Tan 'however'	tz'u^{3} 'this'	liang^{3} 'two'	che^{3} 'entity'
chih^{1} 'of'	hsing^{4} 'sex'	chuang^{4} 'appearance'	chi^{2} 'and'
yung^{2} 'flow'	chieh^{3} 'loosen'	tu^{4} 'measure'	shao^{1} 'somewhat'
wei^{2} 'do'	pu^{4} 'not'	t'ung^{2} 'alike'	

The translation is:

"The appearance and solubility of these two substances, however, are somewhat dissimilar."

Bibliography

John P. Cleave 67
Braille Transcription and Mechanical Translation
Mechanical Translation, Vol.2, No.3, pp.50-53, (Dec. 1955).

This paper presents a comparison of similarities between the problems arising in mechanical translation and the mechanical transcription of texts into Braille. Both processes require the development of formal rules, stated in terms of word patterns in the input text, which prescribe operations to be performed. The routine needed for Braille transcription is much simpler than that required for mechanical translation because of the small vocabulary consisting of only a limited number of letters and punctuation marks, the absence of ambiguity, and the fact that explicit rules, already partially formalized, exist. A summary of these rules is included in the paper.

J.R. Applegate

Alexander Gode 68
The Signal System in Interlingua--A Factor in Mechanical Translation
Mechanical Translation, Vol.2, No.3, pp.55-60, (Dec. 1955).

This paper is an attempt to justify the use of Interlingua, an artificial language whose syntactic categories are those common to a group of base languages, as an intermediate language to be used in mechanical translation. The author states that translating from Interlingua to a base language would be easier than translating between base languages. Translations from Interlingua might be comprehensible without editing.

J.R. Applegate

I.S. Mukhin 69
An Experiment of the Machine Translation of Languages Carried out on the BESM
Academy of Sciences of the USSR, Moscow, 1956, 28 pages.

A short description of the methods used in programming the BESM computer to translate from English to Russian. A vocabulary of 952 English words was used, augmented by a routine for separating the affixes s, 's, ing, ed, er, est, e, th. Special routines were used with 121 of the English words for distinguishing multiple meanings on the basis of the preceding or following words. The program analyzes the English sentence in a similar manner, changes the word order and synthesizes the Russian sentence, adding inflections to stems.

V.H.Y.

Silvio Ceccato 70
La grammatica insegnata alle macchine
Reprinted from Civilta delle Macchine, Nos. l and 2, 1956.

This monograph presents a discussion of the problems encountered by members of the Italian Operational School in their attempts to develop techniques to be used in mechanical translation. The monograph includes sections on language structure, semantics and general problems of translating from one language to another. There is also a discussion of three things which the author considers the three difficulties in developing a program for mechanical translation. These are: the lack of exact correspondence of meanings in the source language and the output language, the fact that one word may have several meanings, and the lack of an exact description of sentences. The article ends with the presentation of a program to be used in translating the English sentence, "The cat fears the dog," into Latin.

J.R. Applegate