[*Mechanical Translation*, vol.4, nos.1 and 2, November 1957; pp. 35-43]

The Thesaurus in Syntax and Semantics†

M. M. Masterman, Cambridge Language Research Unit, Cambridge, England

The recent work of the Unit has been primarily concerned with the employment of thesauri in machine translation. Limited success has been achieved, in punched-card tests, in improving the idiomatic quality and so the intelligibility of an initially unsatisfactory translation, by word-for-word procedures, from Italian into English, by using a program which permitted selection of final equivalents from "heads" in Roget's Thesaurus, i.e. lists of synonyms, near-synonyms and associated words and phrases, instead of from previously determined lists of alternative translations. The Unit is investigating whether the syntactic properties of a word in a source language may be defined by a simple choice program, with reference to extra-linguistic criteria, which might be of universal or extensive interlingual application. It is hoped to combine or reconcile such a program with R.H. Richens's procedure for translating syntax by means of an interlingua, which has proved effective in a small-scale test. Studies have been made of the complementary distribution in literary English of words and phrases from "heads" in Roget, and of the construction of discourse from the contents of selected "heads." The possibility of producing a thesaurus better suited for machine translation purposes than Roget's, to be based on a more restricted lexis and a simpler categorization, is to be examined.

AT THE Second International Conference on Machine Translation, held at the Massachusetts Institute of Technology October 16-20, 1956, members of the Cambridge Language Research Group[1] presented four papers[2] which together opened up a new approach to certain linguistic problems of machine translation. As a result of discussions which followed, a Research Unit was formed at Cambridge, with the support of the National Science Foundation of the United States, to investigate these problems further.[3]

One of the great problems of machine translation is that of providing any device, programmable on a machine, for translating idiomatic or metaphoric uses of word when these uses cannot be foreseen, since they may be occurring for the first time in the language which is being translated. To meet this problem, three of the Cambridge research workers, M.M.Masterman, A.F.Parker-Rhodes and M.A.K.Halliday, recommended that a mechanizable procedure for producing non-literal, "idiomatic" translations should be tried. This procedure required an

---

† This paper has been written with the support of the National Science Foundation, Washington, D.C.

1. The Group is a private, informal research society, most of whose members hold appointments in the University of Cambridge (see MT. Vol. 3, No. 1, p. 4). The Unit, concerned specifically with machine translation and library retrieval methods, was formed mainly from members of the Group, with some additional workers.

2. M.Masterman, "Potentialities of a Mechanical Thesaurus"; A.F. Parker-Rhodes, "An Algebraic Thesaurus"; R. H.Richens, "A General Program for Mechanical Translation between Any Two Languages via an Algebraic Interlingua" (reported MT, Vol.3, No.2); M.A.K. Halliday, "The Linguistic Basis of a Mechanical Thesaurus", now published MT, Vol. 3, No. 3.

3. See Annual Report of the National Science Foundation 1957 (in the press).

extra dictionary, compiled not on the principles of an alphabetic dictionary, but of a thesaurus,[4] to be inserted into the machine handling the target language. Thus, if the target language were English, the main part of the procedure would consist in retranslating an initially unsatisfactory translation, obtained by the word-for-word procedures long known to be feasible in machine translation, into idiomatic English. The actual translation procedure, moreover, did not consist, as had all mechanical translation procedures up to that time, of programing the machine to make a selection between the members of a finite set of antecedently given translations of a source language word. It consisted, on the contrary, of a procedure for mechanically producing from a thesaurus a finite set of extensive lists of synonyms of a particular word; that is, of a total dictionary in miniature; and of then choosing, by a two-stage procedure, firstly from among the lists, and secondly from among the synonyms. Thus, by looking up the word 'plant,' say in the cross-reference dictionary of a thesaurus, a set of numbers can be obtained, each standing for a list of synonyms, which might appear in one context, of the word 'plant:' "plant as place, 184: as insert, 300: as vegetable, 367: as agriculture, 371: as trick, 545: as tools, 633: as property, 780: – 'a battery,' 716: – 'oneself,' 184: – 'ation,' 184, 371, 780." This last represents an actual extract from the cross-reference dictionary of Roget's Thesaurus. Initially, the machine cannot know which of these lists of synonyms of 'plant' it should choose. But suppose that the word 'plant' were preceded, in the text, by the word 'flowering.' The cross-reference dictionary entry for flowering' is as follows: "flower as essence, 5: as produce, 161: as vegetable, 367: as prosper, 734: as beauty, 845: as ornament, 847: as repute, 873: – 'of age,' 131: – 'of flock,' 648: 'of life,' 127: – 'painting,' 556, 559." There is only one context in common between the context list of 'plant' and the context list of 'flowering,' namely, 367, 'Vegetable.' We therefore correctly assume that the synonym list under Vegetable is the synonym list required, if a synonym is in fact required for the basic word 'plant.'

The last stage in the procedure consists in comparing, in twos, the synonym lists which have been selected by the procedure given above in order to find which synonyms occur in common in these. Thus, if 'Woman' and 'Animal' are looked up in Roget's Thesaurus, and the synonym lists under each compared for common words, a single common word will be discovered, namely 'bitch.' These common words are then ordered, in descending order of frequency and the most frequent provide the retranslation output, certain restrictive rules having been brought into play which are designed to decide unambiguously which synonym shall replace each initially given pidgin English word. Sometimes, as in the case of 'plant,' in 'flowering plant,' the output is the same as the initially given word; this is taken as confirmation that the original translation was right. But sometimes, in the test cases presented at the Conference, the final output was significantly different from the original word. Thus, by using what came to be known as the "thesaurus procedure," it was shown that the Italian phrase alcune essenze forestali e fruttiferi. which had been translated, by a word-for-word translation procedure, 'forest and fruit-bearing essences,' could be retranslated 'forest and fruit-bearing examples [or specimens];' that the Italian phrase tale problema si presenta particolarmente interessante, which had been translated, by the word-for-word procedure, "such problems self-present particularly interesting,' could be retranslated 'such problems strike one as, [or prove] particularly interesting;' and that the Italian word germogli, which had been translated by the word-for-word procedure 'sprout,' could, though with difficulty, be retranslated 'shoot.' The papers made clear that the use of such a thesaurus procedure by no means always produced a correct translation. For instance, the phrase particolarmente interessante, which had been correctly translated by the word-for-word procedure 'particularly interesting,' was retranslated by the thesaurus procedure as 'What's the matter?' Nevertheless, the examples showed that a trans-

---

4. The only way of defining the notion of a thesaurus, in practice, is by reference to the famous work of Roget, Thesaurus of English Words and Phrases (Longmans, Green and Co.

5. Locke and Booth, Machine Translation of Languages (New York and London, 1955). See esp. Chapter II; Richens and Booth, Some Methods and Mechanized Translation.

6. I.S.Mukhin, An Experiment in the Machine Translation of Languages Carried out on the B.E.S.M. (Moscow. 1956); examples: 'category' (chart on p. 16); 'of' (chart on p. 17).

lation device which was programable on an electronic digital computer, but which made use of the intrinsic elasticity of words, could hope to deal, in a significant number of cases, with the hitherto unsolved problem of translating idiom, metaphor, and pun.

The fourth paper presented at the Conference, by R. H. Richens, made a different, though cognate, recommendation. In it the author recommended that a completely general interlingual notation, or set of symbols, should be used to produce syntactically correct translations between languages of different types, without any effort being made to translate directly between any given pair of languages. Richens showed, moreover, that by the use of such an interlingua, and by a mechanical procedure so simple that it could be effected not only by a digital computer, but by a punched card machine, a sentence could be translated with complete syntactical correctness from Japanese into the interlingua, and from the interlingua into English, German, Latin and Welsh. Thus the Japanese passage conventionally translated as: KETSU SAKU HO GO HEI ni ICHI SAKU to[2] ri SHU SHI RYU SU[2] ha KO HAI JI KI ni yo tsu te I ru was rendered into English as 'the percentage of matured capsules and the number of grains of seeds of one capsule are different according to the time of hybridizing;' into German as der Prozentsatz der gereiften Kapseln und die Zahl der Grane der Samen einer Kapseln sind gemäss der Zeit des Bastardierens verschieden; into Latin as ratio per centum capsulas maturandi et numerus granorum seminum capsulae unius secundum temporem hybridizandi diversa sunt; and into Welsh as y mae canran oeddfedu masglau a rhif gronynnau hadau un masgl yn wahanol yn ol amser croesi rhywiau. And Richens' claim, made in his paper, that his interlingua was algebraic has since been justified. When subjected to mathematical logical analysis, the Richens interlingual notation was shown to possess the characteristics of a weak mathematical system.

It might be thought that such revolutionary translation proposals as these, requiring as they do such an immense amount of computer storage, would be of merely academic interest to machine translators until computer research had developed to a point considerably in advance of that at which it now is. This is by no means the case, however. Information presented at the same conference, notably in a paper by Dr. Gilbert King, [7] made it clear that in the machine translation field, computer research is far in advance of language research; that, if the linguistic problems can be solved by any mechanizable procedure, computer engineers will find a way of programing the solution on to a machine. At a speech made at the Conference's final day, for instance, Dr. King said that procedures which had been brought forward at the Conference had convinced him that a machine could translate not merely as well as, but better than, an M.I.T. professor; since, having more storage space, it could produce a bigger vocabulary. Thus the papers presented by the Cambridge research workers at the Conference produced an atmosphere of technological hopefulness about the future prospects of mechanical translation, which did not, perhaps, take sufficient account of the fact that the basic linguistic problems, though tackled, were not yet solved.

After the Conference, it rapidly became clear to us that the generality of approach implied by the proposal to use a target language Thesaurus was cognate to, but not identical with, the generality implied by the proposal to use an algebraic syntactic interlingua. The more recent work of the members of the Unit has, therefore, been primarily directed towards making explicit the exact nature of the interrelations between these two proposals. For it is evident, on the one hand, that an interlingual claim is being made by the assertion that Language is such that, in it, metaphors and proverbs can, in some cases, be interchanged by means of a thesaurus. And, on the other hand, the analytic examination of Richens' interlingual algebra has established that it, itself, when interpreted, showed some, though not all the characteristics of a thesaurus. The question therefore arose: could the two methods be unified? Could an interlingual thesaurus somehow be conjoined to an interlingual syntactic notation to produce completely interlingual idiomatic mechanical translation from any language into any other? Conversely, could syntactical correctness as well as semantic elegance be introduced into the translation program at the stage of target-language retranslation by including a syntactic section within a thesaurus, so as to produce idiomatic multilingual mechanical translation from any source language into a single target language ?

---

7. King and Wieselman, Stochastic Methods of Machine Translation (International Telemeter Corporation, 1956).

Up to this point, the nature of the mechanical translation technique had required that the major part of the Cambridge Unit's analytic work should be performed by programmers and mathematical logicians, not by linguists; for the Unit's first need was to produce an analysis of the translation process which was both sufficiently general to justify the commercial production of a future mechanical translator, and also mathematically definite enough to be mechanizable. Now, however, it became clear that essential and fundamental considerations, regarding both the nature of comparative descriptive linguistics, and the nature of philosophic logic, were tied up in all this analytic work. For, to mention only one such consideration, the promoters of the thesaurus target-language procedure could, and on occasion did, claim that they were mathematicizing Plato; Richens, with an equal justice, could be said to be mathematicizing Aristotle. Thus, with sophistications on both sides, the age-old controversy in philosophy between nominalists and realists took, in the research conferences of the Cambridge Language Research Unit, a strange, fascinating, esoteric new turn.

Secondly, it became clear that if a well-grounded decision was to be made between the policy of interlingualizing the thesaurus, (that is, of assimilating semantics to syntax) and that of thesaurizing the syntax (that is, of including syntax within semantics) the linguists would have to be called in. In fact, for a time, they would have to be given charge. In the attempt to decide between these two alternatives, the Unit had developed two complementary lines of research. In the first, Richens designed an interlingual program complete with dictionary for translating syntax, beginning with translation from Italian into English, but subject to continual test by translation from other languages. In this test the object was to see how, with a very rough-and-ready method of translating metaphor and idiom, but with a very advanced and sophisticated method of translating syntax, intelligible translations of scientific texts could be made without using a thesaurus. In the second line of research, transformations were made from thesaurus-heads to texts and then back again within one language, without any procedure being used to translate from one language to another, or to translate syntax. The linguists were then invited to comment on and improve both of these lines, in order to see whether or not they tended to contrast or converge.

Halliday's sophistication of the Richens interlingual syntax translation program was of the following general form. For the general description of it I quote his own words:[8]

".. Translation.. is a form of comparative descriptive linguistics; but whereas translation between a given pair of languages requires only particular (one language) and comparative (in this case transfer, i.e. two languages) description, we envisage it as a requirement of mechanical translation that the program should be applicable to translation among all languages, and therefore we must face the necessity of universal (all languages) description ... Clearly if work was concentrated on a one-one translation field, where only a straight transfer description is required, results might be expected much more quickly. But the whole program might have to be remade for each pair of languages, and [so] it seems preferable to aim at a universal linguistic translation program applicable to translation between any pair of languages.

"This wider aim can only be achieved by a rigorous separation of the particular from the comparative universal range of validity (in MT terminology, of monolingual from interlingual features), and by their separate handling in the program ... The basic problem in the grammar is the setting up of relations among the particular grammatical structures of different languages ... It seems clear that considerable use can be made, both in the dictionary entry and in the operations, of the descriptive distinction between those chunks [separable segments of words[9]] which can be fully identified in the grammatical analysis (i.e. grammatical chunks or 'operators') and those only partially identified in the grammar and requiring further, lexical, information (i.e. lexical chunks or 'arguments'). This is of course an arbitrary distinction made for mechanical translation purposes; it reflects the different fields of application of the grammar and the dictionary in

---

8. From "The Linguistic Basis of Mechanical Translation" (Report for the Eighth International Congress of Linguists, University of Oslo, 1957; in the press).

9. See Richens and Halliday, "Word Decomposition for Machine Translation;" presented to the Georgetown University Eighth Round Table Meeting on Linguistics and Language Studies, April, 1957, and to appear in its Proceedings (in the press).

descriptive linguistics ... Comparative linguistics has the theoretical equipment [for establishing a universal description of syntax] by reference to categories of context grammar; and the systems of context-grammar categories set up for mechanical translation make up a grammatical interlingua such that any single language is capable of comparison with them. This grammatical interlingua .. is not a universal language, which would merely turn the number of languages we have to deal with from n to n + 1, but a set of systems of grammatical relations identified in context grammar, of the type that one sets up for the comparative identification of grammatical categories in descriptive linguistics .. The method [of setting these systems up] which seems at present likely to be most fruitful, and [which] is being tried out on a limited number of languages, (Italian, Chinese, English, Russian and Malay in the first instance), is [first] to establish a rigid operator/argument distinction, and [then] to identify the operators by their placing in a number (provisionally about 60) of two term systems each term being a yes-or-no function, . . The arguments are then classified by reference to grouping of these systems .."

Halliday's method, then, stripped to its essentials, is first to make a monolingual grammar of each language, and then, distinct from this, an interlingual analysis. The monolingual grammar is of the kind normally produced by descriptive linguists, except that it is only for the operators of each language; it is by reference to these operators that the arguments are, later, to be defined. This monolingual grammar can, at a later stage, be mathematically related to the interlingual analysis of these same operators, but is initially sharply to be contrasted with it, since it is to be based on extra-linguistic, not on intra-linguistic context.[10] The interlingual analysis, the making of which is the key to the whole problem, is achieved by the following method. With regard to each operator in question, the analyst asks himself a number of extremely simple questions, questions so simple, in fact, that he can unhesitatingly answer, with regard to them, "Yes," "No," "Both," "Neither" ("Neither" meaning "The question is inapplicable"). For instance, take the French operator la, the function of which, for mechanical translation purposes, is always very difficult to define, since, speaking vaguely, it can serve either as a feminine definite article or as a feminine accusative pronoun. We assume that la has already been monolingually placed within a set of monolingual grammatical systems, including a two-gender system, which apply to French only. We therefore feel free to ask, interlingually, not "Does la belong to any gender system?" because it is notorious that gender systems, as between languages, do not correspond, but, far more simply, "Can la, under any circumstances, tell us anything about sex?" Thus, by this change of question, we are exchanging a reference to the intra-linguistic context, (i.e., that of French) for the far more stable extra-linguistic context, i.e., that of the division of the human race into two sexes. English has no genders, French two, German three, Icelandic six; but Englishmen, Frenchmen, Germans and Icelanders alike all fall into communities consisting of two, and only two, sexes. Thus, with regard to the French operator la, when we ask, "Can it, ever, tell us anything about sex?" we can instantly and unhesitatingly answer, "Yes, it does." Proceeding to the next question, we ask, "Does la apply to animate/inanimate objects?" to which the answer is, "It applies to both." To the next question, "Does la apply to present/non-present time?" the answer is, "Neither; the question is inapplicable." "Does la refer to proximate/distant regions of space?" Answer, "Neither; the question is inapplicable. " (With regard to the French operator là this question could be answered; but not with regard to la), and so on. The heart of the whole method lies in the application of the precise and elegant methods used by contemporary descriptive linguistics to analyze monolingual context grammar (methods which amount in effect to analyzing the older compendium units "verb," "adjective," "noun" and the rest into weaker but more stably definable unitary components from which any required variant of the compendium units can be built up) to analysis of extra-linguistic context also (Halliday; June, 1957). In this latter case the extra-linguistic contexts can be universal ones, and the compendium units are the actual operators themselves. In other words, by taking seriously the analogy which has always been known to exist to some extent between intra-linguistic and extra-linguistic context, and by

---

10. M.A.K.Halliday, "Some Aspects of Systematic Description and Comparison in Grammatical Analysis" (Studies in Linguistic Analysis; Philological Society Special Volume, London, 1957).

treating the first as a straight extension of the second, Halliday has shown that he can achieve, for practical purposes, a non-contentious method of universal grammatical description. (By 'non-contentious' I here mean only, 'a method which will produce the same answers to the same questions when applied to the same operators by different analysts.') Moreover, the preliminary use of this method gives some provisional reason to think that the more complete and comprehensive the series of "Yes/No" questions which are asked (however large it is, the list will be objectively determinable and finite) the more closely the numbers of operators in each language come to approximate to one another. The result, if it is confirmed, will be very useful for mechanical translation, since it means that, with regard to any language, the operator category will be checked and redefined by the interlingual analytic process itself.

Thus Halliday's suggestion for sophisticating Richens' translation program is already of considerable research interest, since it shows that even so initially general and purely logical a research project such as that of Richens can be re-envisaged as arising out of a valid linguistic field. Halliday's suggestion is also hopeful in that preliminary research trials show that it does provide a paradigm, or model, for the rapid construction of operator dictionaries. Thus the Unit has plans to prepare such dictionaries in Italian, Standard Chinese, Cantonese, Malay, Hindi, Russian, Turkish, English, French, and German, these being the languages for which the dictionary makers are readily available. If the method justifies itself, other languages, without too much strain, can be added to these. The second consideration which can be derived from studying Halliday's schema is that he is, in effect, making a syntactical thesaurus. Several of the yes-no questions by which he establishes the components of his categories, for instance, "Does this operator apply to animate/inanimate objects?" "Does this operator assert a fact / give an implication?" "Does this operator indicate completion/non-completion?" "Does this operator indicate duration/non-duration?" could equally well be used as part of a schema for classifying synonyms under given thesaurus-heads. Thus a convergence between the interlingual and thesaurus approaches is detectable here.

What is not yet established, as must be made clear, is whether the additional complexity which Halliday desires to insert into the very simple and elegant translation program of Richens will really improve the quality of the translation produced by it. A test is being devised of the capacities of the original and amended versions to translate prepositional phrases. Meanwhile, another feature has emerged, in that Halliday's amendments to Richens' program have strengthened the case for coding this program to go through the computer by using the very general mathematical system known as lattice theory. (The use of lattice theory for the analysis of language will effect an analysis congruent to the ideas of those linguists who can, in any sustained way, imagine language as a net. On a first approximation, a lattice is an asymmetric net; a finite lattice is a fishing net or hammock, though an asymmetric one; that is, a net with a single top point and bottom point. Such nets are built up from a single asymmetric binary relation, which itself derives, though over some distance of time, from the asymmetric binary relation used by George Boole, and which was suggested to him by the linguistic adjective-noun relation.) Preliminary grounds for using this mathematical system to algorithmize the translation of syntax had already been given in earlier papers by the members of the Unit. [11] Moreover, the fact that the Richens interlingua had already been shown to constitute an algebraic system weaker than lattice theory, though not incongruent with it, increased the ground for re-mathematicizing it by trying on it a mathematical system of the same kind as itself, though of more algorithmic power. And Halliday's analysis, being as it is in terms of dichotomies, (and of systems which can be constructed by successions of dichotomies) straightforwardly uses lattice theory by its very nature. Either, therefore, it must be compressed and coded by initially using this system, or it cannot be compressed and coded at all. Some idea can be gathered, however, of the extent of the complication which Halliday's suggestion introduces into Richens' program from the fact that whereas an entry of 20 bits (20 binary digits) per chunk would have sufficed Richens to translate both meaning and syntax, Halliday's amendment will require an entry of at least 120 bits

---

11. See MT, Vol. 3, No. 1, pp. 2-28 (report on the Colloquium of the C. L. R. Group, August, 1955); and M. Masterman, "The Comparative Analysis of a Chinese Sentence, " (annex to the report, available from the Editor of MT).

per chunk for syntax translation alone. Fortunately, Dr. Gilbert King, who was mentioned earlier, and who now is a member of the Unit's Consultative Committee, considers it feasible, from the engineering point of view, to construct a mechanical translator which will perform lattice operations but not arithmetical ones, and which will allow of chunk entries 1, 000 bits long.[12] For existing computers, however, Halliday's schema would be too complex by far. This should not blind us to its intrinsic interest or to its many potential advantages; but it should be borne in mind by those linguists who are seriously interested in developing machine translation as a concrete reminder that, for every increase in linguistic analytic complexity, a heavy electronic price has to be paid.

Turning now from syntax without semantics to semantics without syntax, a word must be said about the Unit's second research project, namely that of examining the interrelations between texts and their constituent thesaurus-heads without the complicating intervention of a foreign language. Dr. E. W. Bastin, Karen Jones, M. M. Masterman, R.H.Needham, A.F. Parker-Rhodes, A.R.Penny, Dr. R.H.Thouless and W.F. Woolner-Bird have made the principal contributions.

The first provisional discovery made by the members of this research group was that paragraphs of lecture-style discourse could, without difficulty, be constructed by the intuitive use of a minimum number of thesaurus-heads. Thus a paragraph dilating pompously but not vacuously on the present peculiar scientific position of the study of parapsychology was constructed by Dr. Thouless and Margaret Masterman, for thesaurus demonstration purposes, using only four lists of thesaurus synonyms to supply all the argument words. These lists concerned the generic ideas of 'Wonder' (with a cross reference to 'Interest'), 'Science,' 'Parapsychology,' and of a very general topic within which 'Appearance in Thought' contrasted with 'Instantiation in Reality,' the two combined heads forming an antithetic pair. The method by which the paragraph was constructed was suggested by one of the Unit's programmers, Lady Hoskyns. If Interest be A1, Wonder A2, Instantiation in Fact B, Psychical Research C and Science D, then the paragraph constructed by Dr. Thouless can be thesaurized as follows:

" 'Interest' [A1] in 'psychical research' [C] is often 'motivated' [A1] by 'wonder' [A2] at 'phenomena [C] which 'appear to be' [B] 'marvellous' [A2]. The 'sitter' [C] is 'amazed'[A2] at the 'wonderful' [A2] 'results' [D and B] of 'card-guessing experiments' [C] which 'leave him in a state of' [B] 'bewilderment' [A2], 'seeming' [B], as they do, 'to savour of' [B] 'necromancy' [A2]. This 'attitude' [A1] of 'awe' [A2] (or of 'admiration' [A2], as it would earlier 'have been called' [B]) 'produces' [B] a 'fascination' [A2] with the 'subject' [C and D]. The 'new-comer's' [C] 'surprise' [A2] 'leads' [B] often to 'stupefaction' [A2], and the 'research' [D] is 'treated' [D] as a 'sensation' [A2] rather than as a 'serious' [A1] 'branch of science' [C and D]."

Other paragraphs, giving the obituary of an imaginary well-known biologist, an advertisement for a film star, and a denunciation of the British Conservative Party, were similarly constructed. The introduction of a randomizing procedure, with the object of mechanizing the selection of synonyms, caused a paragraph of esoteric theology, and also one denouncing philosophic scepticism, to be a little more irrational than they would otherwise have been, but not very much. Attempts rapidly followed to use this method to construct parody (Thouless and Parker-Rhodes); to simulate essay writing (Woolner-Bird); and to employ it to analyze chapters instead of paragraphs (Needham and Jones). Several facts of considerable interest emerged. One was that, in any kind of writing which builds up into an argument, thesaurus-heads tend to be introduced in powers of two, each topic being introduced concurrently with that to which it primarily contrasts. Another was that the introduction of a new thesaurus topic, in discursive writing, tends to follow a clustering of re-allusions to a single one of the topics which have been introduced earlier, and which are themselves synonymous, in such a way as to force the selection of the new thesaurus-head. This result was reached independently by Woolner-Bird and by Needham and Jones (by analysis of Southern, Cultural Aspects of European Territorial Expansion.) A third fact which emerged was that, if the unit to be analyzed consisted of a chapter, rather than a paragraph (that is, of a piece of discourse with an order of, say, 20 enlarged

12. G.King, The Requirements of Lexical Storage (International Telemeter Corporation, 1957).

thesaurus-heads), a sub-class of these heads, say, 2 or 4, will have vastly more synonyms of themselves occurring in the chapter than will any of the others; so that this sub-class of heads, taken in a prescribed ordering, can be taken as a title for the whole chapter. A fourth fact, of very general interest, was that there are some thesaurus-heads which always have to be constructed to analyze discourse; that is, which occur so constantly that it seems almost impossible to think without them. One of these conveys the very idea of a synonym: "is, constitutes, appears to be, seems to be equatable with, shows itself to be, constitutes the fact that; namely, that is, in other words; could be called, could be treated as, could be considered as; this comes to saying, this comes to the same thing as saying. . " These and their like appear in every text; (including the present report). So do synonyms of the very general generic idea of causation: "causes, promotes, produces, leads to, determines, results in; the result is, the upshot is, in the end, we find that we can say that.. " So do synonyms for the very basic idea of appearing to be one thing, while turning out in fact to be another. (This generic idea precedes nearly every introduction of contrast.) Since these thesaurus topics so constantly occur, it might be argued that their constituent synonyms were functioning as a queerly determined class of syntactical operators, rather than as arguments. Moreover, since, in order to analyze the chapter of a book into its constituent thesaurus-heads, a distinction has to be established, and in a non-contentious manner, between new ideas (formalized by P), qualifiers, to be taken as a single element with what they qualify (formalized by Q's) and re-allusions to ideas previously mentioned (formalized by R's); and as all these have to be distinguished from Q's, or operators, it becomes clear that if Halliday, to translate syntax, has to construct a new type of universalized thesaurus, so also the thesaurus makers, in order to analyze the semantic patterns occurring in texts, have to construct a very basic, simple kind of syntax. All of which gives reason to hope that in some way (the members of the Unit do not yet see how) the interlingual program for translating syntax, and the analytic program for constructing texts from thesaurus-heads, or thesaurus-heads from texts, may all turn out to be different parts of the same program, in the end.

In conclusion, a final word must be added on one problem of thesaurus construction which the members of the Unit will have to face squarely if they are to construct a full-scale translation thesaurus. The creative ability of man is not so easily amenable to mechanization, in this field, as the Unit's early, gaily-reached results, would seem to imply. In other words, with every text we analyze it becomes increasingly evident that every discursive writer constructs his own thesaurus. How then is the Unit to construct a thesaurus which has any hope of applying to more than one text?

One immediate reply to this capital difficulty is by asking another question: "How, equally, does any linguist compile a dictionary which fully applies to more than one text?" In a paper on categorization of lexis, recently read to a meeting of the Language Research Group at Cambridge, R. A.Crossland suggested that a procedure of selection out of a thesaurus-head, alternative or preferably supplementary to any procedure based on contextual distribution, might be based on the traditional dictionary-maker's technique of classifying words as appropriate to particular general contexts or types of diction. [13] Such indication is given only sporadically and somewhat unsystematically in most existing dictionaries, but, with refinement, it might provide a technique for programing the computer to make an appropriate choice from among the possible alternatives in a thesaurus-head, especially when this is to be used in the final stage of translation. Two methods of providing this selection suggest themselves. Either information about the appurtenance of a word in a source language to different dictions ("high" or "low" style, the styles of various technologies, etc. [14]), is recorded and passed through the interlingual stage, though the computer in that stage translates just an approximate lexical equivalent (the key word of a thesaurus-head, perhaps). Or else, without the recording and transmission of such information, an appropriate equivalent, out of a head "labelled" according to the appurtenance of its constituent elements to different dictions, would be selected in accordance with general

13. Diction seems now to be virtually a synonym in philological discussion for "verbal or written style" (cf. Oxford English Dictionary).

14. Crossland noted the element of subjectivity involved in categorization not based on detailed analysis of contextual distribution within restricted textual material.

and immediate context, (either by the procedure described earlier, or by some other mechanizable procedure to be substituted for it), within the set of such heads constituting the "rough output."

If any of these suggestions proves fruitful, it would seem likely, on the face of it, that new thesauri will have to be prepared, or existing ones reorganized by "labelling" of items and no doubt by addition, deletion and rearrangement, for languages between which translation is envisaged. Also it might be useful to prepare thesauri on the basis of particular scientific or other specialized "dictions." These could be considered valid in practice for fairly extensive categories of writers, though in principle the argument that every writer has his own thesaurus, based on what he alone desires to write or has written, seems reasonable enough.

Whether the Cambridge Research Unit will really succeed in compiling such a gigantic, universally valid, thesaurus of thesauri is not yet clear. What is clear, in the sense that it is becoming established as a thesis supported by considerable factual evidence, is that when a human being thinks discursively he does use a thesaurus. Secondly, it is intuitively clear, in the sense that it follows from this, that somehow or other, human beings do succeed, in discursive argument, in communicating to one another the boundaries of their respective thesauri; for if they did not, there would be no argument. We know this; for when communication fails to take place, we say, "I cannot understand the writer; he is too allusive." What we say, in making such a comment, is the opposite of what we actually mean; because what we mean is that such a writer does not take the trouble to order and display the re-allusions to his main ideas sufficiently for us to "catch" his personal procedure of synonym creation; that is, sufficiently for us to ascertain his thesaurus. And when we say this, it is further intuitively clear that we must be referring to some objective communication-promoting procedure; some procedure which we use, without being aware that we use it, whenever we argue discursively with one another.

The task that confronts us, then, though formidable, is not hopeless. Objective synonym-creating procedures which can be employed, can also be discovered; and logicians, dictionary makers and descriptive linguists are just the men to discover them.