A NEW RESEARCH TECHNIQUE FOR ANALYSING LANGUAGE

by
E.W. Bastin and R.M. Needham
May, 1959

General Remarks, by E. W. Bastin.

One is very familiar with the situation in a science that is in the stage before final results are available, in which discussion which constitutes cogent argument when viewed with the outlook of members of one broadly discernable school of thought, appears to members of an opposing school as merely a series of isolable speculations. Such a set of apparently isolated ideas are presented by the publications of the Cambridge Language Research Unit on Language.

Bar-Hillel, in his recently published Report*, admits that he is "amazed by the prolificy" of these different ideas, and alleges that he has entertained some of them at various times himself. Nevertheless he regards the various mathematical approaches that have been adopted by the Cambridge Language Research Unit precisely as a series of isolable speculations. He suggests that the Cambridge Language Research Unit constantly pick up one theoretical approach which holds a limited sway before being discarded in favour of a fresh one. He thus puts himself in the position of a member of a school working with an opposing theory to that of the school of thought which he is discussing, in that he is unwilling to entertain the basic presuppositions of any theory alternative to his own.

The conclusion I draw from the fact that Bar-Hillel takes this position is that he despairs of attaining any objective standard (or series of experimental tests) to judge between the rival approaches; and, having come to this conclusion, I find myself deeply critical of Bar-Hillel's understanding of the characteristic mode of progress of a scientific theory. I consider that he should be prepared to judge the value of this progress of successive examination of different theoretical facets of a new approach, retrospectively in terms of its outcome. And its outcome should be an experimentally applicable technique. Thus, I maintain that we do not have

indefinitely to remain in the situation where rival theories conflict, and in which the choice of which we accept depends on which basic presuppositions we entertain. A vital stage in the development of a science is passed when the combination of theoretical discussion and close examination of the experimental material have issue in the discovery of a technique. Moreover, it is the possession of a technique which makes a theory public property - a thing which can be transferred from one group of workers to another without change of meaning and which makes experimental tests about whose significance there can be general agreement, possible.

It is possible to express my criticism of Bar-Hillel - namely, that he does not appreciate the importance of possessing a technique - in a different form. Bar-Hillel's publications - and indeed his criticisms of the various M.T. units - give little evidence that he attaches any importance to that to-and-fro process of trial and error and consequent progressive reformulation of ideas about an unknown subject-matter, that characterises an experimental science. On the contrary, he gives the impression that he regards the problem of M.T. as a complex problem of coding which - in its essentials - the necessary knowledge about the language is already at hand. If, with this approach, Bar-Hillel feels little sympathy with the method of the Cambridge Language Research Unit in their search for a technique, there should be little cause for surprise. The day-to-day work of the Cambridge Language Research Unit depends upon a vital curiosity about language. In general, that work is most highly valued in which some empirical discovery is made which could not have been foreseen at all if the experimental work had not been done. This approach is in sharp opposition to one which assumes that, in essentials, the knowledge about the language essential to Mechanical Translation is already to hand. The C.L.R.U. need their technique precisely because of their essentially empirical approach: it is an agreed technique alone that can give repeatable connection with the empirical material when it is a new theory which is in question.

To conclude, I believe that the Cambridge Language Research Unit has arrived at a stage where it possesses such a technique, and an attempt to justify this will be made in the following remarks, by a detailed reference to their work. I think that Bar-Hillel's failure to understand the importance of a technique has had the result that he has failed to give the C.L.R.U. credit for having successfully reached this stage.

Detailed Remarks, by R. M. Needham.

Before discussing the course of the work of the Cambridge Language Research Unit over the past 18 months, in detail, I shall elaborate upon what I mean by developing a technique. A close analogy is intended with the process of setting-up apparatus for a series of experiments in a physical science. The setting-up process often takes a very long time, and in the course of it, a great many alternatives are tried and rejected. From the outside, it may appear that much time is wasted in futile tests and in trying systems that have subsequently to be rejected. However, the experimenter has his aim constantly in mind, and only if he is a very bad experimenter, does he allow his basic theories on the matter in hand to be obscured by the ebb and flow of practice and discussion at the early stages. This I wish to show has been exactly the situation of the Cambridge Language Research Unit; I refer to the quite concrete task of discovering how to put into practice and try out the Unit's ideas.

At the beginning of the period under discussion (about November, 1957), the Unit had to begin experiments on the application of the thesaurus-intersection procedure as described by M. Masterman (1), and refinements of a way of looking at syntax through sets of questions due originally to M.A.K. Halliday.

Accordingly, three experiments were carried out by hand. That is to say, all the logical operations involved were performed by people looking at lists of things. Two of these concerned the syntax questions ("Magnam Multitudinem Vidit" and "Ad Ludum Ambulamus"), and endeavoured to carry out an intersection procedure on the syntax questions between the various parts of the Latin words (2), The third was concerned with the semantic program ("Agricola Incurvo...", attached), and arrived at a syntactically unrefined, but semantically plausible translation output.

It here became clear that in the future some kind of mechanisation would be absolutely necessary as the clerical labour of performing the tests by hand was very great, and the results were very unreliable. Accordingly, a decision was taken, which has influenced the whole subsequent course of the Unit's work. This was to use punched-card (i) machinery as far as possible rather than a computer. The reasons for this unusual

---

(i) Hollerith cards had already been used for the Library Scheme (3)

course were as follows:

1. In the testing stage of procedures, punched-card machinery is likely to be much quicker. This is because, although the machinery itself is much slower than a computer, the programming is much easier and quicker. This is clearly economic in time if tests are in question which are liable to be very much altered after a little experiment. Thus, it is better to spend two days setting up punched-card machinery, one day running the programme, and then deciding to alter it radically, than it is to spend two weeks programming a computer, ten minutes running the test, and then deciding to alter.

2. It is much easier for all of the Unit's research personnel to be personally capable of organising and carrying out tests for punched-card machines than for a computer (i).

Further points on this head appear below at their appropriate place in the chronological account.

When this decision to use punched-card machines had been taken, the British Tabulating Machine Co.(ii)was approached; they very kindly agreed to lend the Unit machinery free of charge and also to supply cards. The first machine was a specially modified duplicating punch which would, from two cards, prepare a third with only the holes in common between the first two, regardless of the number of holes punched per column. This machine enables experiments to be performed which involve the Boolean meet and join operations upon cards. The reason for requiring this operation is that it figures prominently in the tests mentioned above, and proved very

---

(i) Y.H. Yngve's autocode COMIT (4)is designed to meet the same research need as caused C.L.R.U. to use Hollerith machines. The relative success of the methods cannot be judged yet; it appears to me that COMIT's strength will lie in re-ordering and substituting techniques and that of the Hollerith machinery in logical flexibility.

(ii) Now International Computers and Tabulators Ltd.

awkward to do reliably by hand. Although comparison of cards can be done by a "Peek-a-Boo" method, repeated comparison requires mechanical reproduction of the result card.

When this machine arrived, work was started on more full and elaborate tests of the same procedures that had been employed previously (2) (5). This involved the preparation of a coding scheme for representing the thesaurus heads of a word on a card. It was decided to reserve the top rows of the card for syntactic information, and to try to represent the Roget heads of a word on the remaining 800 holes. Because of the head intersection procedure to be used (5), it was desirable to have each head represented by one of the 800 positions. However, there are 1,000 heads in Roget's Thesaurus, so some measure of compression, or abbreviation, was needed. After investigation (6), it proved possible, without loss of information, to effect the reduction, and accordingly the heads used for experiments thereafter were the 800 heads of what was called the "Compacted Roget".

It was then decided to conduct a test of the syntactic and semantic procedures on a section from the first book of Caesar's Gallic War. This was the famous passage, "Gallia est omnis divisa in partis tres, Quarum....". When this was started, it rapidly became apparent that the syntactic procedures (as in (2) were quite inadequate, and would not work at all. The semantic procedure, as in ( ) worked satisfactorily. However, much greater care had been taken with the dictionary entries than for the case of "Agricola...", and it turned out that the set of heads remaining as the "translation specification" was, in each case, considerably larger. If the procedure in (5) is recalled, it is clear that this results in a very tedious and unreliable manual operation at the last stage, consisting of examining a number of Roget heads for common words. It became clear that the next effort of mechanisation must be made on this stage.

This resulted in the beginning of work on a "fan-card dictionary" for English (7). This consisted of a set of cards for English words or groups of related English words, with their heads in the Compacted Roget (see above) punched in the same hole per head code. To discover common words

between heads, it would be sufficient to pass the pack through a sorter sorting successively on the holes representing the required heads. (The Hollerith sorter had by this time arrived.) This process is a definite, though tedious, one. While the fan-card pack was being made, the heads given in Roget for the words were scrutinised and many necessary additions made.

Concurrently with this work on the fan-cards, an idea was explored which might have use in dealing with the syntax. It was essentially a return to the much older idea (1) (8). To mechanise this process as far as possible, methods were devised for using the machinery at hand (9). These devices, in particular, one for using the duplicating punch to find the most fluent heads of the passage, should be very useful in other applications.

At about this time advice was sought from punched-card specialists (10) on the general techniques being used and to be used. Conversations with them resulted in a programme being drawn up (11) which has since been known as the Staniforth programme. It brought to the Unit's attention the essential point about punched-cards which is that it is possible to have indefinitely large packs of them. One can dispense completely with limitations of space; for experimental purposes, it is possible largely to dispense with those of time. The opinion of these consultants reinforced the belief that the decision to use punched-card methods was the right one, particularly as the present programmes operated in effect with 800-bit words, which is very awkward to do on most computers.

Various improvements on the Staniforth programme and refinements of procedure led to the preparation of the "Library for M.T." (9), a collection of punched-card procedures.

The work described above took up most of 1958 up to September. Concurrently with it a great deal of work was done on improvements to the Thesaurus. All the thesaurus-using tests to date had been based on Roget's Thesaurus (Penguin Edition), and had shown up ever more clearly its defects. It is inconsistent, clerically incomplete, contains too little vocabulary, and is rather obsolete. For these reasons, the progress

of the fan dictionary was slow, since it was as much a work of lexicography as of transcription . When the basic library of programs was brought to the stage of the description, it was decided that further expansion and tests must await the availability of sufficiently large scale dictionary material. This in turn depended on improvements to the thesaurus, to which the Unit's attention was then turned. While the details of this are not strictly a matter for the present paper, it falls within the province of technique to discuss the general attitude of the C.L.R.U. to lexicographic work.

A consequence of the Unit's theoretic approach is that very high-quality dictionaries are needed. This means that they cannot be made quickly (12). The general principle has thus been to make only as much of a dictionary as is necessary to perform the tests contemplated, since it will probably turn out that the content, or format, of the entries will be inappropriate for future work. There are two dangers in this course; firstly, that the entries made will be "cheats", and secondly, that no large scale work will ever be done. The first can only be avoided by caution. The second has in practice been avoided. The Unit has constructed an Italian-Nude dictionary for the testing of R.H. Richens' translation scheme (13), not described in this paper, of some 7,000 chunk entries with a translating power of well over 20,000 words. The fan dictionary runs to some 600 cards covering about 4 English words each; work on it is being resumed now that the Italian dictionary is complete.

The foregoing account has been intended to show how the efforts to carry out in practice the theoretic ideas of the Unit as at November 1957 have resulted in the acquisition by the Unit of a great deal of necessary practical knowledge, and how in the process of doing this points of theoretical importance have come up. It is not intended as a complete account of the activities of the Unit members. Particular omissions are the work of sorting out and collating the Halliday questions before the first experiment(14); the theoretical aspects of the lexicographic work on Latin (16); computer experiments on a bracketting programme (17); work on the data-processing aspects of the ultimate computer translation method (15); and the extension and testing of the library scheme (18).

Subsequently to the course of work which culminated in the "Library for M.T.", the Unit has received a collator, and is to receive a reproducing punch, with the aid of which experiments on Richens' programme become more possible. To the previous body of techniques are being added others which apply to the syntactic part of Richens' or similar methods (19). Currently work is going on to determine the best division between the kind of methods of the "Library for M.T." which are basically mechanised peek-a-boo methods, and more normal punched-card methods. The operative point is that while the peek-a-boo methods are exceedingly efficient for particular operations, they result in the use of multiple-punched-cards (i.e. more than one hole per column) which is a severe drawback for other operations. A final decision on this is not yet available. Probably a certain amount of duplication will be accepted for practical convenience; that is, some packs will exist in both multiple-punched and Hollerith-punched forms.

E.W. Bastin
Research Fellow of
King's College

R. M. Needham
Cambridge Language Research Unit
Cambridge University Mathematical Laboratory

References:

1) Masterman, M., Potentialities of a Mechanical Thesaurus, MIT Conference on M.T. 1956.

2)"Magnam Multitudinem Vidit". C. L. R. U. Workpaper.

3) Joyce, T., Needham, R.M., The Thesaurus approach to Information Retrieval, Amer. Doc. 1958.

4) Yngve, V. H., The Comit System for M.T.. UNESCO Conference, Paris, 1959.

5. "Agricola Incurvo.....". Appended to this paper.

6. Shaw, Fr. M., Compacting Roget's Thesaurus, C.L.R.U. Workpaper.

7. Jones, K.S., Blackmore, R.M., "Fan Cards", C.L.R.U. Workpaper, unpublished.

8. Parker-Rhodes, A.F., Appendix to (1) above.

9. Jones, K.S., Blackmore, R., Wordley, C., Library for M.T. using punched-card machinery, C.L.R.U. Workpaper, attached.

10. They were R. A.Fairthorne, S. Whelan, J. Staniforth.

11. Staniforth, J.M., General Schema of a Thesauric Translation Programme Using a Punched-card Technique. C.L.R.U. Workpaper, attached.

12. Masterman, M., Dictionary Entry for the Latin "IN". C.L.R.U. unpublished.

13. Richens, R.H., The Thirteen Steps. C.L.R.U. Workpaper.

" " , Interlingual M.T.. Computer Journal, 1958.

14. Jones, K.S., C.L.R.U., unpublished.

15. Parker-Rhodes, A.F. & Needham, R.M., Encoding Roget's Thesaurus. C.L.R.U. Workpaper.

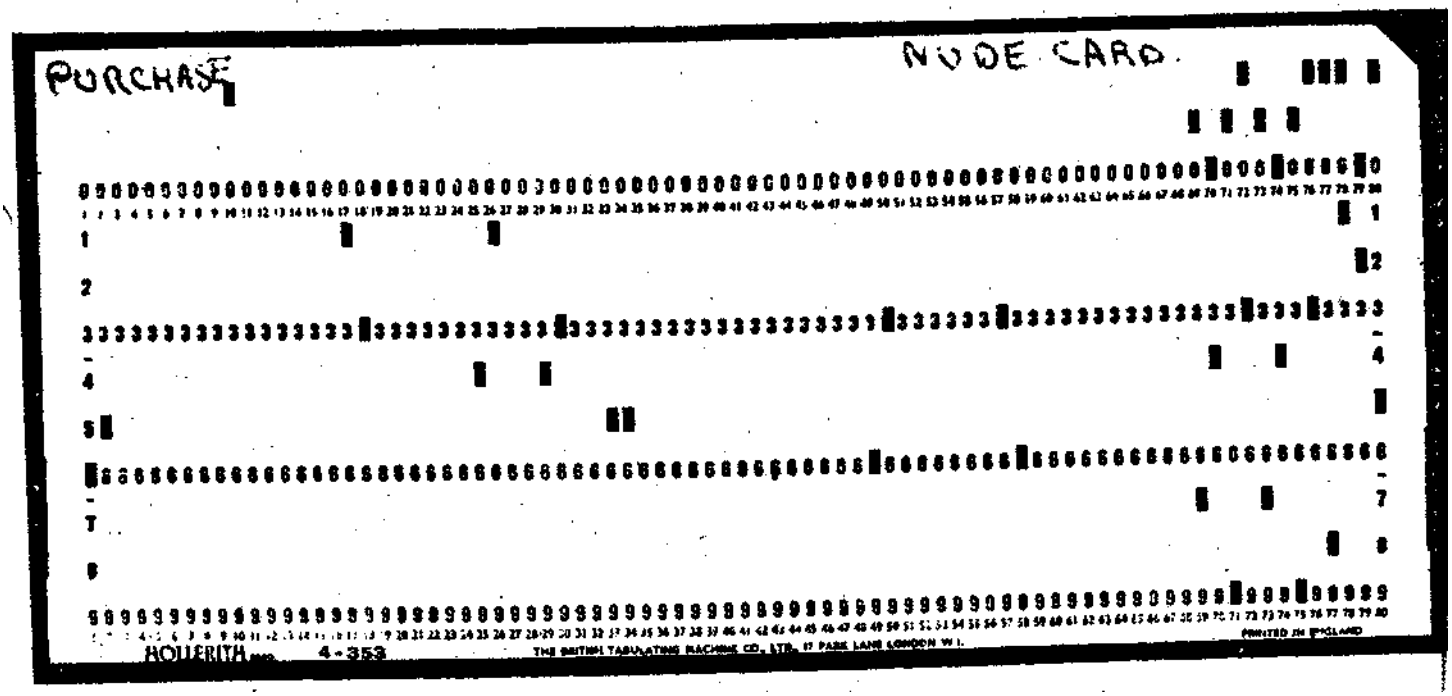
16. Masterman, M., C.L.R.U., unpublished.

17. Parker-Rhodes, A.F., C.L.R.U., unpublished.

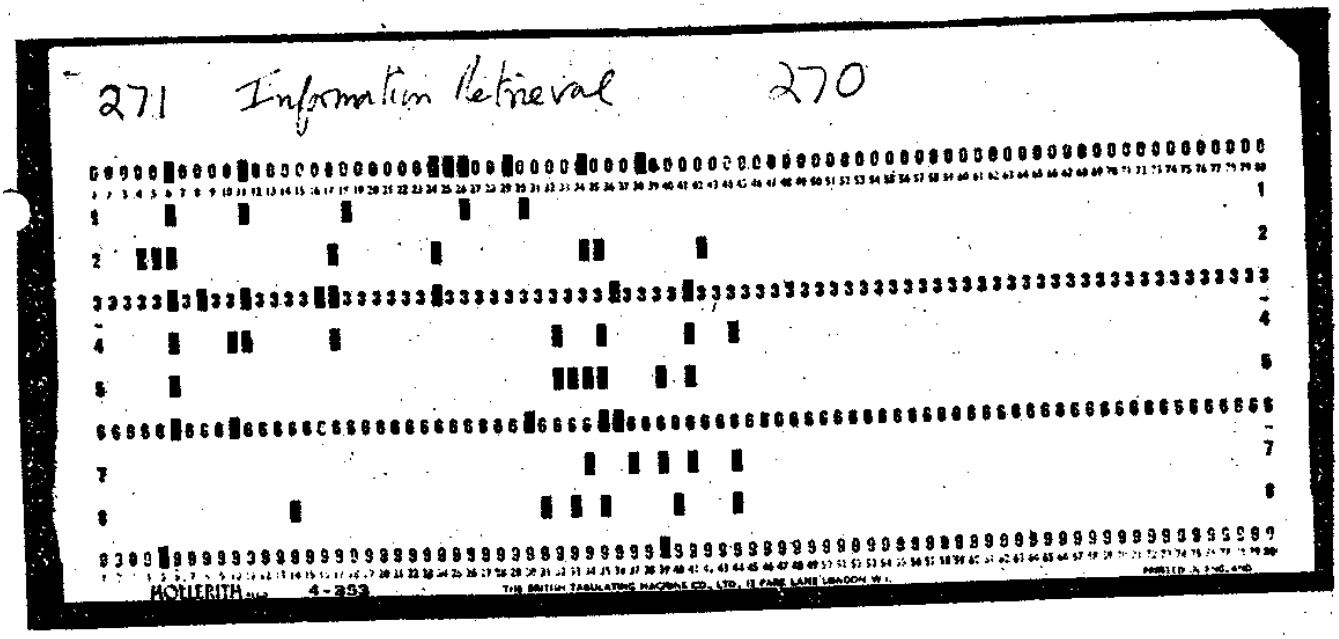
18. Miller, A.H.J., Extension & Testing of C.L.R.U. Library Retrieval System. C.L.R.U. Workpaper.

19. Kay, M., Marcode - a description of a method of encoding the formulae of the Interlingua "Nude" on punched-cards, C.L.R.U. Workpaper.

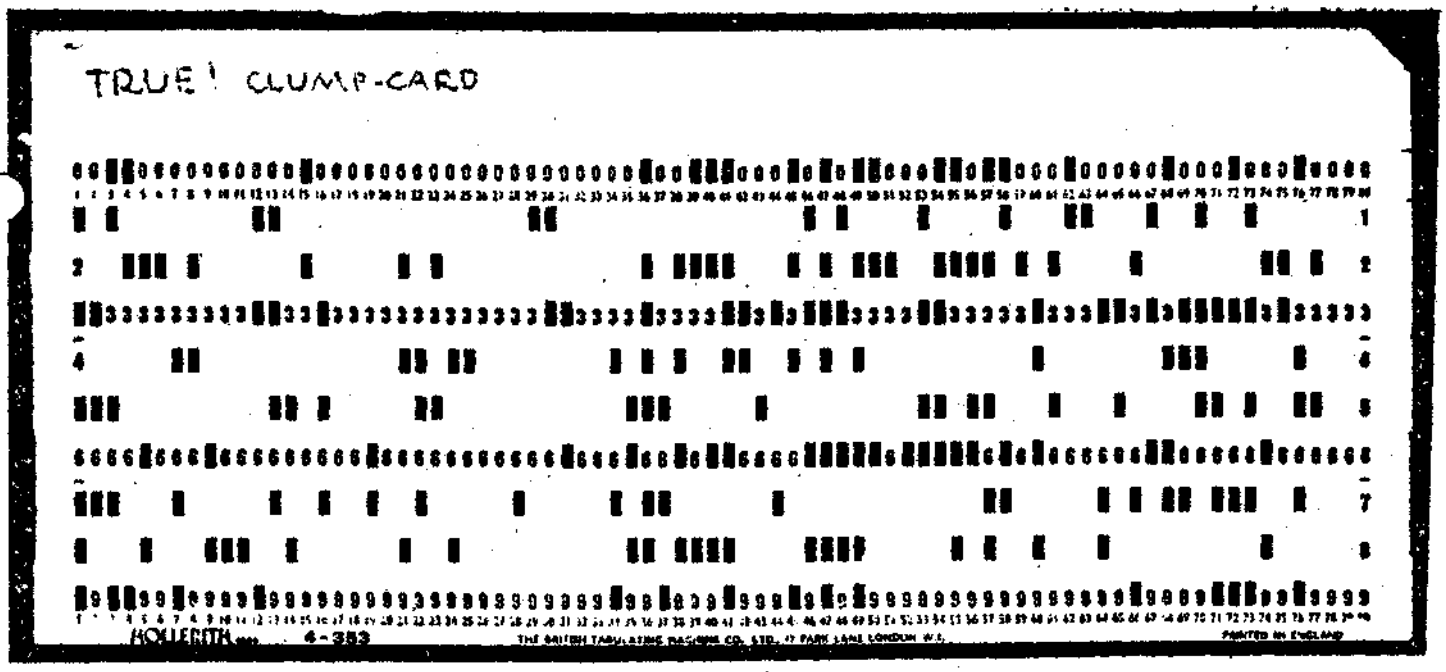
SPECIMEN NUDE CARD



SPECIMEN INFORMATION RETRIEVAL CARD



SPECIMEN CLUMP CARD



SPECIMEN INPUT DICTIONARY CARD

(CHUNK OF AN IMAGINARY LANGUAGE)