[*International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, Teddington, UK, 5-8 September 1961]

MECHANISED SEMANTIC CLASSIFICATION*

by

K. SPARCK JONES

(Cambridge Language Research Unit, England)

INTRODUCTION

IT is now widely admitted (see, for instance, de Grolier (1)) that a semantic classification will be required for machine translation and information retrieval; and that as mechanised procedures will be carried out on it, it must be detailed, precise, and explicit. This paper is primarily concerned with the construction of such a dictionary, rather than its use, i.e. with applied language analysis as a preliminary for machine translation.

Apart from the problem of finding a suitable form of classification, the labour of compiling a dictionary of this kind is very great, and mechanisation of some, if not all, of the drudgery involved is desirable. The need to tackle the whole question has become more urgent, for it has become clear that reasonably high quality machine translation requires a higher standard of dictionary making, and in particular a more detailed, i.e. more realistic, representation of the full range of uses of a word than has hitherto been considered necessary. This is brought out, for example, by the inadequacies of the IBM output which is obtained on a word-for-word basis (2).

As a solution to the problem of providing a refined but manipulable classification the Cambridge Language Research Unit has advocated the use of a thesaurus (3,4,5), i.e. a system of conceptual groupings. To construct such a classification, therefore we must

i) give a workable procedure for carrying out the extremely refined linguistic analysis required for a complete treatment of the

* This paper was written with the support of the United States Office of Naval Research, Washington D.C.

word-uses of a natural language; (this is emphasised by the defects of existing thesauri such as Roget (6) ;)#

ii) give criteria for obtaining conceptual groupings from this material.

It is clearly desirable that the methods adopted should be as objective as possible. While I do not pretend that the procedure given for carrying out the initial analysis is mechanisable, the subjective element is minimised, and the results are thoroughly suited to machine handling. Once this initial analysis has been made, however, the conceptual groupings are obtained by wholly mechanical means.

In the system described below the initial analysis gives classes or "rows" of synonymous word-uses, i.e. word-uses which are mutually replaceable in at least one linguistic context. (For the purposes of the classification the specification of word-uses in terms of their synonymity relations is regarded as adequate.) By using the hypothesis that word-uses with the same sign are in general more like than those with different signs, second-order classes can be obtained representing concentrations of common signs over sets of rows, i.e. representing semantic closeness in sets of rows, i.e. conceptual groupings. Computer experiments on English are then described.

1. PRIMARY CLASSIFICATION

The first object of this investigation is to find a way of defining* a word-use which is both semantically adequate and a suitable basis for further classification; i.e. we are looking for an appropriate form of mechanisable dictionary entry.

The simplest approach, i.e. that of going direct to the extra-linguistic reference, (at present being studied by M. Masterman) has the disadvantage that difficulties about "the mechanism of reference" immediately arise $^{\delta}$. If however, we look at the way in which a word is used in a sentence, the referential problems need no longer concern us: for although they ultimately arise when the relation of the whole sentence to

---

Text-scanning has been suggested as a solution to this problem. If treated merely as a device for obtaining examples of word-uses, however, it has to be carried out on a very large scale if adequate coverage is to be obtained; and the resulting material, such as that collected for prepositions by Yngve (7), has still to be classified. The suggestion has also been made that the classification itself may be carried out on the basis of the co-occurrence of words in sentences obtained in this way. But the information required can only be obtained in an even more dilute form than the preceding, and I know of no suggestions for turning this vague idea into a practicable procedure.

* Except in the formal system "definition" is used in the sense of "specification"

δ See for example, Quine (8).

its reference is considered, we can, if we assume that the sentence is understood, disregard them. This approach is essentially that of linguistic philosophers such as Austin (9) who show how a word is used by giving examples of the kinds of linguistic contexts in which it can occur. The method as it stands is merely illustrative, and therefore unsatisfactory because the resulting samples of text cannot themselves be mechanically handled.* I shall show, however, a) that we can make use of this sort of information without having to give it in full, and b) that the relevant facts about the way in which a word is used can be "encoded" in a suitably compact and tractable form.

The formal system

1. A ***sentence*** *is* a finite sequence of elements (words), bounded by terminal characters, having a property called a ***ploy*** (the way in which it is employed).
2. A sentence may have more than one ploy.
3. The same ploy may be common to two or more sentences.#

The ***length*** of a sentence is the number of elements which it contains.

Consider the class S_i of sentences specified as having the ploy P_j. We will assume that this class has more than one member. Consider the sub-class Σ_i of S_i containing all the sentences in S_i having a particular length L_m. We again assume that this class has more than one member.

Let σ_i be the sub-class of Σ_i, again of more than one member, such that:

1) the element at a particular position k in each sentence in σ_i differs from that occurring at the corresponding position k in every other sentence in σ_i;
2) the element at every other position in each sentence in σ_i is the same as that occurring at the corresponding position in every other sentence in σ_i.

The elements a,b,c, ... occurring at k in the sentences in σ_i will be said to be ***parallel*** with respect to k in σ_i.

4. A class of elements which are parallel with respect to some position n in some class σ_n will be called a ***row***.

We can thus, for every position n and every class σ_n obtain a row; for a particular class σ_i we can obtain a row for every position; and for any sub-class of a class σ_i we can obtain a row for each position

---

* This is also true of Aristotelian definitions in which the extra-linguistic reference is described.

\# For practical reasons we shall consider written texts only.

which will be different from that obtained for the same position by σ_i itself or by any of its other sub-classes.

If we make a pairwise comparison between the members of a class σ_i, we can say for each pair that at the position k where the elements differ, the element in one of them has been replaced by the element in the other; the two are otherwise, both formally (i.e. in length etc.) and in ploy, the same. We can plausibly and to practical advantage therefore say that we are dealing with one sentence and a class of elements which can replace one another at a particular position in it without changing its ploy. As the members of a row are thus mutually replaceable, the class of elements constituting a row will, as before, be finite and unordered.*

A revised definition of row can now be given:

5. A finite class of elements will be called a row if its members are mutually replaceable with respect to a position n in a sentence s_n.

We have so far used the expressions "element" and "word". The aim of the system, however, is to deal with word-uses, not with words, and it is also clear that in starting from sentences we are in fact concerned with word-uses and not words, in the classification, moreover, individual word-uses are treated as separate units. If, therefore, we are to give meaning to "use of a word", which in the introduction we loosely equated with "word use", we must define "word" in terms of word-uses.

A sentence was defined as a ployed sequence of words. We should more strictly have said "sequence of word-signs representing word-uses"; i.e., a word-sign represents a word-use because it occurs in a ployed sentence. Our basic assumption that words are best defined in terms of their uses means that the most appropriate definition of a word will be as the class of its uses, i.e. as the class of uses with the same sign. We now formally define "word-use" and "word" as follows:

6. A ***word-use*** is the occurrence of a word-sign in a (ployed) sentence.

7. A ***word*** is the class of occurrences of one word-sign.

Thus a sentence (c.f. Defn. 1) is both a sequence of word-signs and a sequence of word-uses, and a row is both a class of word-uses and a class of word-signs (cf. Defn. 5).

---

* "Replacement" is used rather than "substitution" to emphasise the fact that although the element is changed, the ploy is preserved.

Comments on the definitions.

1) We say that we can define a word-use by listing synonymous uses*; this gives us, as required, a definition which is obtained intra-linguistically and which is unstructured, concise, and complete. For although the system is developed in terms of classes of uses, it is clear that as the members of a row are equally synonymous, each member is specified by the class of remaining members. As definitions in proper form of individual uses can thus always be given, there is no harm in taking the classes as our units when further classification is required, particularly as the practical advantages of doing this are obvious.

2) The fact that we are dealing with word-uses and not words means that we can construct a classification based on synonymity which is nevertheless far more flexible and far more realistic than the usual logicians' total synonymity will allow#. For it must be emphasised that although the replacement criterion is extremely strict, it need only hold in one case, and its range is therefore extremely limited. We can moreover obtain empirical support for the assertion that our approach is a satisfactory one by reference to standard dictionaries: the entries in the (large) Oxford English Dictionary, for example, often consist of sets of synonyms or near-synonyms which are very like our rows; a typical instance is "CIVIL : humane, gentle, kind". (Other entries can without significant loss of information be reduced to this form: thus, for example, "CLOD: a coherent mass or lump of any solid matter, e.g. of earth, loam, etc." would

* In assuming that it will be clear whether two or more uses are synonymous, i.e. that on replacement the ploy of the sentence remains unchanged, we can only rely on the linguistic judgement of the dictionary maker. This may seem inadequate, but we can argue that a subjective element must enter all lexicography at some point, and that here the point at which it enters is carefully defined, and the scope which it is allowed is extremely limited.

\# The logicians' interpretation of synonymity as "a can always be substituted for b" (10) is connected with discussions of logical truth, analyticity etc., and has therefore a specialised purpose. It must be pointed out however, that these discussions make use of examples from ordinary language, where synonymity in this sense is rare, and are to this extent dangerous.

The real nature of synonymity in natural languages is recognised, on the other hand, by A. Naess (11); he allows synonymity between two word-uses each of which occurs only once. He is mainly concerned, however, with setting up procedures for testing synonymity in particular cases, and makes no attempt to base a general classification on the "synonymity-facts" which he finds.

become "CLOD": mass, lump"*) In actually constructing a classification for English as described in SECTION III, therefore, we can plausibly make use of the Dictionary, either taking the entries as they stand, or using them as the basis of a more elaborate classification still. The fact that we can thus utilise, in the most straightforward way, the very detailed and highly documented information contained in the O.E.D. is important; for although it has been frequently observed that the Dictionary is a valuable source of linguistic information, no suggestions have hitherto been made as to how to encode this material in mechanisable form.

II SECONDARY CLASSIFICATION

By applying the procedure described above we can, in principle, obtain a row for every position in every sentence. Although we may not go to this extreme, and although we do not distinguish identical rows derived from different sources#, it is clear that carrying through an analysis of this kind on a large scale will result in the creation of a very great number of rows. (There will clearly be far more rows than words.) However, as our object at this stage is adequate definition and distinction, the degree of refinement represented by the procedure is an advantage; for the multiplicity of rows directly reflects the multiplicity of distinctions made in the language, and if high-quality machine translation is to be achieved, we cannot afford to ignore such a basic feature of language. Nevertheless, if the classification so far constructed is to be really useful, we must derive from these first-order classes a much smaller number of second-order classes; and the latter must, if the system is to be thesauric in character, in some sense represent conceptual groupings. (By conceptual groupings we mean, to put it crudely, groups of rows which refer to similar extra-linguistic situations δψ.) These classes must, moreover, be

---

* Some O.E.D. definitions are "irreducible" descriptions of the Aristotelian type; in our system this means that the words are not replaceable, i.e. are undefined, such words, or "technical terms", do not, however, represent a breakdown in the system: for they are intended to be, to an unusual extent, precise in reference and unambiguous in use, and synonyms are excluded to avoid the possibility of confusion, or because they would be redundant. Technical terms cannot in fact be adequately handled in any purely intra-linguistic classification, and must be given special treatment. It should be noted, on the other hand, that, in contrast to "ordinary" words, they rarely present problems in translation.

\# Rows which are sub-rows of other rows are kept separate.

δ They need not be, and almost certainly will not be, mutually exclusive.

ψ For formalisation of the notion of extra-linguistic situations see Masterman (12).

obtainable intra-linguistically by objective and mechanisable means, or the first-stage restrictions on subjectivity and intuition will be wasted. Yet the only intra-linguistic information available for connecting rows is that represented by words, i.e. the recurrence of common signs: and from this we cannot on the face of it, deduce anything about the semantic relations of the rows or even, indeed, of the semantic relations between the uses of a word.

However, if we look at a collection of rows, groups which overlap in containing common signs nevertheless strike one as representing conceptual groupings. We shall suggest that this is a consequence of ***the fundamental fact that, in a language, there is a finite number of signs for a much larger, and constantly expanding, set of situations, and that if this were not so, effective communication in ordinary circumstances would be impossible.***

In a given sentence-position the members of a row are, by definition, mutually replaceable: i.e. there is a choice among the different members of the row. We shall say that this choice is one between different signs for a particular "Word-use". The point of this interpretation is that the Word-use is determined by the relevant extra-linguistic situation, although the choice of signs is not. The Word-use, therefore, in contrast to the signs, is genuinely interlingual, and, when we communicate, is what we want to get across. Thus the second-order classes we require will be genuinely interlingual classes of Word-uses.

The reasons why we can derive conceptual groupings from overlapping signs are best understood if we first consider what happens in other kinds of language.

In a language in which a Word-use is represented by a single arbitrary sign, such as a technical language, or a code like the International Code of Signals, there is no intra-linguistic information, not even recurrent signs, on which to base conceptual groupings. The latter can only be obtained by considering the situations to which the signs refer, i.e. by subjective and extra-linguistic means. A conceptual grouping, moreover, can only be specified by listing its members: there is no intra-linguistic aid to remembering the relations between them. A language like this is indeed worthwhile only where unambiguity is more important than convenience, and only usable if it is comparatively small and used in well-defined circumstances. Given such a language with a much larger number of situation references, it is clear that conceptual groupings can only be handled if an economy in the number of signs is somehow effected.

To achieve this economy we might

(1) use the same sign for very distinct Word-uses. As the latter are not semantically related, however, we can only interpret the sign by listing

the uses. The economy is not, therefore, a very helpful one. Moreover, conceptual groupings can only be obtained, as before, by going outside the language.

ii) use the same sign for similar Word-uses (any ambiguity is thus almost harmless); i.e. we can treat a sign as a "shorthand" for a set of similar references. We could also use this information to pick up conceptual groupings, for we know that the members of a set of Word-uses with one sign are semantically related. (The groupings themselves will be more easy to handle, for the number of signs will be smaller than the number of references.) The extent to which we can build up conceptual groupings in this intra-linguistic way is, however, limited: for we can only group sets of Word-uses by considering the relations between the corresponding sets of extra-linguistic situations. Moreover, Word-uses with the same sign can only be distinguished by external reference.

Bearing these points about languages in which there is only one sign for a Word-use in mind, we can now consider a language of the kind dealt with by our primary classification in which Word-uses correspond to classes of word-uses i.e. in which there are both several signs for each situation-references and also a very large number of situation-references.

We can clearly argue that since such a system represents a natural language, and that such a language must, if it is not to be unusable, economise on signs, each word is a shorthand for word-uses with similar references*: i.e. ***we make the Fundamental Assumption that it is in general true that word-uses represented by the same signs are semantically close***. The fact that a particular sign is used for certain word-uses is thus not arbitrary, and we can give a semantic interpretation to the definition of "word"#. This situation is clearly like the one described above, in which we used one sign for several similar Word-uses. That was, however, unsatisfactory: firstly, because Word-uses with the same sign could not be distinguished intra-linguistically; and secondly, because there were no intra-linguistic connections between the sets of Word-uses although the sets themselves could be intra-linguistically obtained.

In contrast, the system represented by our primary classification does not suffer from these disadvantages. For if, and we have assumed that this is both possible and normal in natural languages, we specify word-uses by others, the uses of a particular word are distinguished by the differences in membership of the rows in which they occur; i.e. the distinctive character of a word is represented by the particular class of Word-uses (rows) into which its uses fall, and each of these Word-uses is specified

---

* or groups of similar references.

\# I am excluding the case here of genuinely fortuitous homonyms between word-signs in the language.

by the particular class of word-uses which make up the row. Moreover, ***the fact that we are dealing with combinations of word-uses makes it possible to specify likeness between Word-uses, and therefore to obtain conceptual groupings, by wholly intra-linguistic means:*** for as the members of a row are by definition synonymous, i.e. semantically the same, and as each word-use in a row is connected through its sign to other uses which are by our Assumption semantically similar, we can pick up semantic connections between a row and others which do not all contain the same sign. We are thus not limited to the class of Word-uses with a particular sign, but can link a Word-use with different signs to the different classes of Word-uses associated with each sign in the original; i.e. from a Word-use with sign a we can only go to others with a, but if we start with an a and b, we can go to others with a and others with b.

It is clear, however, that semantic connections depending on one sign alone will not be strong enough to give us very satisfactory conceptual groupings: for although we have assumed that the need for economy forces us to use the same sign for similar word-uses - I will call this the Economy Device - we cannot deduce from this anything very definite about the degree of similarity between the uses. We know, at most, that in general these uses will be more like than those represented by different signs. In classifying rows on this basis, therefore, we can only infer that rows linked by the same sign are more likely to refer to similar situations than those without any common signs; and if the connection (provided it exists at all), between pairs of rows in a potential group is of this weak kind, the group as a whole will not be a very "coherent" one.

But although the Economy Device is in any particular case a somewhat weak semantic tool, if it is generalised we can use it to better advantage: for we can draw the conclusion that ***the greater the proportion of common signs, the more alike two Word-uses will be;*** *i*.e. that if a,b,c and d, members of row A, are synonymous, and a,b,c and e, members of row B, are synonymous, and a qua member of A is probably like a qua member of B, b qua member of A probably like b qua member of B, c qua member of A probably like c qua member of B, this strongly suggests that although d and e are different, the Word-uses of A and B are very similar. We are thus saying that although it may be an accident that one sign occurs in each of two rows, it can hardly be an accident that several do. (This argument is reinforced by common sense).

By using these multiple overlaps, therefore, it is clear that we can obtain genuine conceptual groupings, and, moreover, by wholly intra-linguistic and mechanisable means, i.e. by operations on the signs alone. For the general conclusion about the similarity of pairs of rows can be used as the starting point from which definitions of similarity over sets

of rows can be developed

In order to carry out concrete experiments on these lines we thus require:

i) a precise measure of the similarity of a pair of rows;

ii) a precise criterion of the degree of similarity which must hold over a set of rows if it is to be regarded as a conceptual grouping.

A large number of alternative measures and criteria can be constructed. The measures and criteria actually used in the experiments described below were chosen because programmes based on them already existed. They are taken from work on classification called the theory of clumps by A.F. Parker-Rhodes and R.M. Needham, and will only be described in sufficient detail to make the experiments clear. For further information see the Cambridge Language Research Unit progress reports by Parker-Rhodes and Needham.

The similarity function for a pair of rows was:

$$S = \frac{\text{Number of word-signs in common}}{\text{Total number of different word-signs}}$$

This definition is due to T. T. Tanimoto (13).

The grouping or "clump" criteria were:

i) B-Clump

A set C is a clump if

a) for all $x \in C$, $y \notin C$, $S_{xy} < \theta$ where θ is a suitable threshold.

b) C is maximal for this property (the whole set being excepted).

ii) Kuhns' Clump

A set C is a clump if

a) for all x,y, $x \in C$, $y \in C$, $S_{x,y} > \theta$;

b) there is no $C' \supset C$ such that C' satisfies a).

This definition is due to J.L. Kuhns of the Ramo-Wooldridge Corporation.

iii) GR-Clump

The definition of GR-Clump makes use of the notion of "bias"; the bias of an element x to a set s is

$$b(x,s) = \sum_{y \in s} s_{xy} - \sum_{z \notin s} s_{xz}.$$

A set C is a clump if

a) all members of C have positive bias to C;

b) all non-members of C have negative bias to C.

This definition is due to R.M. Needham.

III EXPERIMENTS TO DATE

The experiments are still in progress and only tentative conclusions can be drawn.

1. Experimental Data

The data for the experiments was obtained from the Oxford English Dictionary. From the point of view of obtaining reliable results from the experiments the problem was that of giving a set of rows which would be both a fair sample linguistically and small enough for reasonably efficient computing. It was decided that the best solution to the linguistic difficulty was as follows:
a small number (approximately 20) of words, some with a wide range of uses, some with a narrow, but each having uses in common with some of the others, was selected; a set of rows for the ***whole*** range of uses of each of these was then worked out as in the example given below. The total set obtained therefore included a number of heavily overlapping rows, others having only one word in common, and some with no common elements. (No completely independent rows were included.)

The rows could not always be "lifted" straight from the O.E.D. as will be seen from the example below; some knowledgeable interpretation on the part of the dictionary maker was required, and the result can therefore be criticised on this ground. But it can be seen that the rows obtained are unlikely to be wrong, though they may be inadequate: and more rows can be inserted if required. The important point is that if the Dictionary is accepted as a "concentrate" of English texts, the results obtained can reasonably be regarded as having a proper empirical basis.

The following is a sample "transformation":

OED: Task. I.1. *A* fixed payment to a king, lord, or feudal superior; any impost, tax; tribute. *Obs.*
e.g. Blackstone's *Commentaries*: "By Statute 25 Edw. I c 5 & 6 ... it was enacted that the king should take no aids or tasks but by the common consent of the realm".

2. A piece-of-work imposed, exacted or undertaken as a duty or the like. Originally, a fixed or specified quantity of labour imposed on or exacted from a person; later, the work, appointed or assigned to one as a definite duty.
e.g. Johnson's *Idler:* "She .... appoints them a task of needlework".
b. spec. A portion of study imposed by a teacher; a lesson to be learned or prepared. Now *arch.*
e.g. Franklin's *Essays:* "These lesson might be given every night as tasks".

3. In more general sense: Any piece of work that has to be done; something that one has to do (usually Involving labour or difficulty); a matter of difficulty, a "piece of work". Froude's *History of England:* "He had taken upon himself a task beyond the ordinary strength of man".

(II, Phrases, III, attrib. and comb.)

Rows: TASK IMPOST TAX

TASK DUTY

TASK JOB PIECE-OF-WORK

TASK WORK

TASK LABOUR

TASK CHORE

TASK OCCUPATION

TASK LESSON HOMEWORK

Some points should be noted. Phrases, etc. (II and III) were omitted for experimental purposes, chiefly because they presented coding problems. (They do not present any real theoretical problems.) The very misleading descriptions "*arch.* and *obs*". were disregarded. In this example the OED entries were not very row-like: the best example is "impost, tax" under 1. In those cases where some interpretation has been required, it must be remembered that information about other words can legitimately, and indeed, should, be used: for a row defines all its members equally, and although the ones given have been listed with TASK first for convenience, they could be given with their members in any order.

The following is a sample of the total set of rows obtained:

PERFORMANCE ACTION WORKING OPERATION

BUSINESS OCCUPATION PROFESSION

ACTION PLOT

ACT STATUTE

MOVEMENT ACTION MOTION

BRISKNESS BUSINESS SMARTNESS

OPERATION WORKING

These rows could be regarded as satisfactory in that the degree of refinement was uniform, that the uses of some words were exhaustively classified, and that the interconnection between the rows over the whole set could be taken as representative. For the experiments a subset of 180 having the same properties as the initial set was selected. (It is intended

as a control, to use more than one subset for the experiments, which will differ in, for example, degree of "inbreeding", average length of rows, etc)

2. Clump-finding

The experiments were carried out on EDSAC II, the Cambridge University Mathematical Laboratory Computer, as part of the research into the theory of clumps. It is expected that with present techniques experiments can be carried out on up to 1000 rows; work, is in progress on more powerful methods for handling larger quantities of data.

A similarity matrix using the function given was computed.
For example with

row 1 ACT DOING
row 2 ACT PERFORMANCE WORKING OPERATION

the entry S_{12} would be 1/5.

The order of the criteria corresponds to the difficulty of finding groups which satisfy them. B-Clumps are mutually exclusive, and though Kuhns' Clumps are not exclusive, there are so many of them that they may not effect any reduction in the data (that is, there may be more of them than there are rows). GR-Clumps do not appear to suffer from these defects; but they cannot at the moment be found in a large set without a lead on where to look. B-Clumps and Kuhns' Clumps, which can be so used, are also, in such a new field of classification, interesting in themselves.

a) *finding B-Clumps*

A search was made for B-Clumps with thresholds θ = .062 (.062) .496*. At the last point many of the individual rows were isolated. There were also 8 small groups, 1 large one, and approximately 70 single rows. This was not very satisfactory; since all similarities, except that of a row to itself, are less than 1, it is obvious that the total set must break up as the threshold is increased, eventually into single elements. Because there is no a priori way of determining a suitable threshold, B-Clumps can only be regarded as significant if they all appear together at a particular increase in the threshold. This condition did not hold for the clumps found. Thus, although the clumps found looked fairly sensible, there was no indication of whether they were the only ones which could have been found. One would only expect to find B-Clumps with material of this kind if, to take an extreme example, it consisted of sets of rows dealing with subjects as disparate as nuclear physics and egyptology.

---

* i.e. roughly 1/16 (1/16) 1/2; the step was slightly diminished for computing reasons.

b) finding Kuhns' Clumps

A search was made with various thresholds: .2, .25, .3 and .34. The latter threshold was very high as it excluded any clump containing more than one two-member row, and in fact gave very few and rather small clumps. The clumps obtained for .25 appeared sensible: a number of these were what could be described as various versions of essentially the same clump. There was a very large number of clumps.

The reasons for proceeding from these two kinds of clump to GR-Clumps will now be apparent:

1) both B- and Kuhns' Clumps depend heavily on a threshold which cannot be other than arbitrary;

ii) there were too many Kuhns' Clumps.

As mentioned above, however, a lead is required for finding GR-Clumps, and this could be provided by using the larger Kuhns' Clumps as "seeds".

c) finding GR-Clumps

In this experiment 7 "seeds" were used giving rise to 4 quite different satisfactorily large clumps (4 of the seeds led to exactly the same clump). All but one of them proved to consist of rows all containing the same one common word, though not necessarily all the rows containing this word: this is a natural consequence of using a small sample, and of using a sample which was obtained in the way described (i.e. starting with a small number of words and finding all their rows): for a large number of rows will not be "pulled" in any other direction because their remaining members do not occur elsewhere.

Specimen GR-Clump

156 TASK LABOUR
163 LABOUR EXERTION
165 LABOUR WORK
186 LABOUR ACTIVITY
167 LABOUR TROUBLE PAINS
168 LABOUR SERVICE
169 LABOUR PRODUCTION
170 LABOUR EMPLOYEES
171 LABOUR PROLETARIAT WORKING-CLASS
172 LABOUR SOCIALIST
173 LABOUR TRAVAIL
174 LABOUR WORKERS
176 TOIL LABOUR TRAVAIL

177 TOIL EFFORT LABOUR TRAVAIL

179 TOIL TASK LABOUR WORK

In this specimen LABOUR and TOIL act as the focal points of the clump, though not all the rows containing LABOUR appear in the clump; thus 77 BUSINESS LABOUR EXERTION PAINS TROUBLE and 80 BUSINESS LABOUR are members of a clump centred on BUSINESS, and 164 LABOUR PERFORMANCE figures in a clump centred on PERFORMANCE.

III CONCLUSION

Granted that the only tests at present available for whether mechanically generated clumps are "correct" are intuitive ones, the results of the experiments were satisfactory. We have thus shown that by using mechanical aids it may be possible to obtain, in a precise and self-consistent way, the kinds of semantic classification required for machine translation and information retrieval.

Note on further experiments

It is thought that present techniques will be suitable for finding clumps in systems of up to 1000 Items, and much larger experiments will accordingly be carried out as soon as possible. (As noted above, a variety of different samples will be used in these experiments).

As a practical matter, it is much easier to find not clumps of rows based on common words, but clumps of words based on common rows. There is a clear duality between the two procedures: i.e. they will extract the same information. If this alternative approach was adopted, a different definition of similarity would perhaps be more natural than the present one:
Suppose we take the ratio:

$$\frac{\text{Number of rows containing a pair of elements a,b}}{\text{Number of rows containing a}}.$$

This is clearly the conditional probability that given that a word a is appropriate in a particular sentence-position, we could replace it by b. This is unsuitable as it stands because it is asymmetrical, but we may conveniently substitute as the similarity of a and b the geometric mean of the two probabilities:

$$\frac{\text{Number of rows containing a and b}}{\sqrt{\text{Number of rows containing a}} \times \sqrt{\text{Number of rows containing b}}}$$

Although experience suggests that the results obtained are not heavily influenced by the choice (within reason) of similarity function, a function such as the one just given which has an obvious interpretation in

the system to which it is to be applied will clearly be more suitable. Further investigation of this question will be one line of future work.

REFERENCES

1. de GROLIER, E. "Tendances actuelles en matière de classifications et codifications documentaires", 7e Rapport au Comité du Federation Internationale de Documentation, Milan, 1960.
2. Output privately communicated.
3. MASTERMAN, M., "The Potentialities of a Mechanical Thesaurus", International Conference on Mechanical Translation, M.I.T., 1956.
4. MASTERMAN, M., NEEDHAM, R.M., and SPARCK JONES, K., "The Analogy between Machine Translation and Library Retrieval", International Conference on Scientific Information, Washington D.C., 1958.
5. MASTERMAN, M., "Translation", to be read at the Joint meeting of the Aristotelian Society and the Mind Association, Cambridge, 1961.
6. ROGET, P.M. "Thesaurus of English Words and Phrases", Penguin Books, London, 1953.
7. Output privately communicated.
8. QUINE, W.V., "Word and Object", New York, 1960.
9. AUSTIN, J.L., "A Plea for Excuses", Presidential Address, ***Proceedings of the Aristotelian Society***, 1956-7.
10. QUINE, W.V., "From a Logical Point of View", Cambridge, Mass., 1953.
11. NAESS, A., "Interpretation and Preciseness", Oslo, 1953.
12. MASTERMAN, M., "Translation", *op.cit.*
13. TANIMOTO, T.T., "An Elementary Mathematical Theory of Classification and Prediction", I.B.M. Corporation, Yorktown Heights, New York, 1958.