[Mechanical Translation and Computational Linguistics, vol.8, nos.3 and 4, June and October 1965]

Experiments in Semantic Classification

by K. Sparck Jones, Cambridge Language Research Unit, Cambridge, England

It is argued that a thesaurus, or semantic classification, may be required in the resolution of multiple meaning for machine translation and allied purposes. The problem of constructing a thesaurus is then considered; this involves a method for defining the meanings or uses of words, and a procedure for classifying them. It is suggested that word uses may be defined in terms of their "semantic relations" with other words, and that the classification may be based on these relations; the paper then shows how the uses of words may be defined by synonyms to give "rows" or sets of synonymous word uses, which can then be grouped by their common words, to give thesauric classes. A discussion of the role of synonymy in language is followed by an examination of the way in which multiple meaning may be resolved by the use of a thesaurus of the kind described.

The work described below has arisen from the Cambridge Language Research Unit's original ideas about the use of a thesaurus for machine translation.[1] Their argument, put simply, was that most words (and not just some awkward words) have ranges of uses, or, as it is sometimes put, have different meanings, or express different ideas, on different occasions. In discourse, any individual word considered by itself is thus potentially ambiguous because it can be used in different ways. This ambiguity is resolved, and the correct use of each word specified, by the surrounding context. This is because a piece of discourse is concerned with, or expresses, a particular idea or set of related ideas. Discourse does not consist of a sequence of semantically unconnected sentences (it would be very hard to understand if it did), but of sentences in which the same key concepts are repeated. The appropriate uses of ambiguous words are therefore picked out because they express the idea or ideas that recur; or, to put it the other way round, the recurring idea or ideas specify the appropriate uses of ambiguous words. The argument is therefore that discourse is essentially repetitive, because without repetition there would be too much ambiguity.

This argument may be correct, but it is too vague as it stands; for machine translation something more definite is required. It was therefore suggested that a precise model of this situation could be constructed by the use of a thesaurus, as follows: words in a thesaurus are classified under different conceptual headings corresponding to the ideas that the words may express; thus, if a word has different uses, this fact will be represented by the occurrence of the word, along with any synonyms or near-synonyms, in a number of sections under different headings. The words in a particular section, or "head," will thus form a conceptual grouping of some kind. If we are dealing with discourse, and we suppose that the words concerned have been thesaurically classified, we can resolve ambiguity by looking for recurring heads. That is, we replace the words in a piece of discourse by the sets of heads defining the uses of each word, and we carry out a set-intersection procedure.

Small-scale experiments on this basis were carried out in the C.L.R.U., using an existing thesaurus, the Penguin edition of the Roget's *Thesaurus of English Words and Phrases,*[2] published by Longmans. These experiments were only moderately successful, and it was clear that this was due mainly to the defects of the *Thesaurus.* A number of words did not occur in it at all, and others were under-classified, that is, they were not listed in enough heads to distinguish all their uses. As it seemed that most existing thesauri would be inadequate for the purpose of machine translation, the question of constructing a better thesaurus, specifically for machine translation, was considered. This would involve

i) better analysis of word uses
ii) checking the headings.

The Problems of Thesaurus Construction

Much of the thesaurus research that has been carried out in the C.L.R.U. has been concerned with the second problem, namely, with the investigation of Roget's headings, and with the construction of alternative sets of such semantic "classifiers"[3]. This approach, however, suffers from the disadvantage that there is always a danger of the headings being *a priori;* we can always ask whether any particular headings are the right ones, and there may be no very obvious way of deciding whether they are or not. A further and more serious difficulty is that it may not be at all clear whether the classification based on a set of headings will have the properties we desire. I have therefore concentrated on the problem of finding a method of constructing a thesaurus in which the *a priori* element is reduced to a minimum.

We can look at a thesaurus head in two different ways: either as a set of words that all come under one heading, or as a set of words that are semantically related to one another in some way, usually as synonyms or near-synonyms.* Of course, if a set of words all come under one heading, they must be semantically related, and if a number of words are semantically related to one another, they will come together under some heading. But the difference between these two ways of looking at a head can help us in considering how we may construct a thesaurus. If we look at a head as a set of words that are semantically related, we are concentrating on the relations between the words in the head, rather than on the relations between the words and the heading. The point about looking at a head in this way is that it suggests that we may be able to construct a thesaurus by analysing word uses in such a way that we pick up the synonymy and near-synonymy information on which groupings can be based. By doing this, we may be able both to obtain an efficient analysis of word uses, and to avoid the difficulties that arise with *a priori* classifiers. There is a further important practical consequence: for anybody actually engaged in making a thesaurus, the ease with which he can decide whether a particular word should be placed in a particular head matters, and it may well be easier to decide that a word should be placed in a particular head because it is synonymous with the words already there, than that it should be placed in the head because it somehow "expresses the notion that the heading stands for."

What we require, therefore, are

1. a method of identifying word uses, to give us our initial data;
2. a method of grouping word uses, to give us our thesaurus heads.

These two procedures must, moreover, give us the refined, precise and machine-usable semantic classification that we require for machine translation.

The Specification of Word Uses

Definitions of word meanings can be either linguistic or extralinguistic. We can sometimes give an extra-linguistic definition of a word, for example by pointing at the thing it stands for, or by giving a picture of it. For our purpose, however, extra-linguistic definitions, even where they can be given, are both unmanageable and inadequate;† there is no very obvious way of storing physical objects in a computer, and many words, like 'resentment' or 'infinity', for instance, have no clear-cut physical reference. Pictures present the same kind of problem. So the kind of definition we use must be a linguistic one. Linguistic definitions can take various forms. One is descriptive: "scowl: a distortion of the forehead, especially a deepening of the lines between the eyebrows, indicating concentration, determination, opposition or hostility." Definitions of this kind are again not easily handled in machine operations. Their variety in structure, length, and level of detail means that they cannot, for instance, be readily compared. Another form of definition is implicit rather than explicit. This is where the meaning of a word is illustrated by exhibiting its use in contexts. The use of 'frown' may be illustrated, for example, as follows: "When she told her father about Mrs. Blenkinsop's visit he frowned, and then said 'I don't think Mrs. Blenkinsop is a very desirable friend for you'." But this kind of linguistic definition is as unmanageable as the first; there is no easy way of picking up similarity and dissimilarity in contexts. A third possibility is to define a word by giving other words with the same meaning or use, that is, to give synonyms, as, for example, in "anger: irritation, annoyance, vexation." This kind of definition, unlike the others, can be coded and handled without difficulty; there are no real problems in sorting and comparing word lists. Moreover, the fact that people, and many dictionaries, such as the *Oxford English Dictionary* (O.E.D.),[4] do define the meanings of words in this way suggests that this is a satisfactory method.

The point about this form of definition is that we are not defining a word directly, in the sense of analysing or explaining its meaning, but rather indirectly, in terms of its synonymy relations with other words. We are saying that 'A' in some sense means the same as 'B', rather than that 'A' means B. We can say that this form of definition distinguishes the intra-linguistic meaning of a word, as represented by its relations with other words in the vocabulary, from its extra-linguistic meaning or reference (in the widest sense of 'reference'), though this distinction is to some extent a matter of emphasis; to put it crudely, we might say that 'poverty' and 'indigence', for example, are synonymous because poverty and indigence are the same state. We are not, therefore, saying that the synonymy relations of a word give everything about its meaning, or that its extra-linguistic reference is irrelevant; the latter is obviously relevant to our understanding of a language. We can nevertheless assume that we know the extra-linguistic reference of a word, so that we can concentrate on its intra-linguistic meaning, since a definition of a word in terms of its synonymy relations may be adequate for our purposes.

In giving a synonym definition, we are making use of a more general idea, namely, that of defining the intra-linguistic meaning of a word in terms of its relations with other words, where these relations may not simply

* There are other kinds of head in Roget's *Thesaurus,* such as the subject groupings exemplified by 267 NAVIGATION, which contains all the words for anything connected with navigation, but the synonym type of head is much more common, and can be regarded as characteristic.

† The question of what kinds of words can have extra-linguistic definitions is thus quite irrelevant to the present purpose.

be synonymy relations, but may include other such "semantic relations." It may indeed be that synonymy is neither the only, nor the most appropriate, relation we can use for defining 'meaning'; and we should now, therefore, briefly consider the question of defining meaning in terms of other semantic relations.

The Definition of Intra-Linguistic Meaning in Terms of Semantic Relations

For our purpose we need a manageable, straightforward relation or set of relations. Dictionary-making depends on the language-user or native informant, so we want to make the procedure for establishing whether two words are related in a given way or not as unambiguous and simple as possible, and this requires well and clearly defined relations. From this point of view, an obvious approach is to use substitution frames in some way. There are a number of relations that might be called semantic relations, and several have been discussed in some detail. The idea that the meanings of words are determined not merely by their reference, but by their place in the vocabulary, and that the vocabulary of a language has a structure, has indeed been developed by linguists following de Saussure and Trier, but little attempt has been made, other than by Lyons, to define the relations involved. (For a survey of this field, see Ullmann, *Semantics*[5].) This is not the place for a full-scale discussion of this subject, so we shall only give some examples of possible semantic relations:

1. *association* (Bally)[8]
   'boeuf' fait penser à 'vache, taureau, veau, cornes, ruminer, beugler . . .'
   'labour, joug, charrue . . .'
2. *hyponymy* (Lyons)[7]
   'tulip' is a hyponym of 'flower', in that "tulip" implies (in some suitable pragmatic sense of 'implies') "flower," but "flower" does not imply "tulip."
3. *antonymy* (exemplified by antonym dictionaries, Lyons) from Smith's *Complete Collection of Synonyms and Antonyms*[8]*:* 'befriend' has as antonyms 'oppose, discountenance, thwart, withstand . . .';
   according to Lyons, 'married' and 'single' are antonyms, in that "not married" implies "single" and "married" implies "not single."
4. *incompatibility (*Lyons)
   'red' and 'blue' are incompatible, in that "red" implies "not blue," but "not blue" does not imply "red."
5. *collocation* (Firth)[9]
   "boy" goes with "sings," but "mountain" does not go with "sings."
6. *synonymy (*exemplified by synonym dictionaries)
   from *Webster's Dictionary of Synonyms*[10]*:* 'dark' has as synonyms 'dim, dusky, dusk, darkling, obscure, . . .'

There are other possible relations, but the problems that arise can be discussed in connection with these. The difficulties are:

i) are they genuine semantic relations?
ii) are they operationally definable?
iii) are they linguistically important?

The trouble with some relations, for instance collocation, is that they bring up the fundamental difficulty of deciding whether a relation is a semantic, that is, linguistic, relation or not. Does the relation between "boy" and "sings," for example, reflect the meaning of the words 'boy' and 'sings' or extra-linguistic facts? We indeed become involved at this point in such questions as whether the statement "The mountains are singing," is a contingent falsehood or something else (a "category mistake"). The philosophical bog that surrounds these questions suggests that it may be difficult to come to any conclusion, but we have to make a decision if we are to proceed with our practical purpose, and it can be argued that in such cases we are dealing with physical rather than linguistic facts, and therefore that this kind of relation is not a genuine semantic relation. Other relations, such as association and hyponymy, turn out not to be satisfactorily definable, or at least not definable in such a way that rapid and non-contentious dictionary making can depend on them. There seems to be no way of giving rules for determining whether one word "makes one think" of another or not, and there are similar difficulties in defining the pragmatic implication that is required for hyponymy or incompatibility. One can see that "tulip" implies "flower" in some obvious sense, but if one starts with, say, "goodness" or "similarity" or "container," the implied terms are less obvious. With "tulip" and "flower," moreover, the implication really depends on the existence of a class-inclusion relation that is doubtfully linguistic. Lyons asserts that hyponymy, incompatibility and antonymy are fundamental to language, but does not give any justification for this assertion, and as it seems, as we have indicated above, that hyponymy and incompatibility cannot be defined satisfactorily, there is no way of discovering whether this assertion is correct. Antonymy could perhaps be defined, not in terms of implication, which is unworkable, but by substitution which reverses the sense of the text in which the substitution is carried out, though this suffers from the disadvantage that it is often hard to decide whether the substitution really does give the reverse or opposite sense.

The general conclusion, therefore, is that most of the potential semantic relations are either not genuine, or not definable. I hope to show, however, that synonymy is both genuine and definable, and, moreover, that it is the fundamental relation determining the vocabulary structure of a language. This means both that we can use synonymy to give us our definitions, and that these definitions will be adequate as specifications of the meanings of words.

The Definition of Synonymy

Synonymy, unlike the other semantic relations, has been extensively discussed, chiefly by philosophers and logicians; and Carnap's approach in *Meaning and*

Necessity[11] represents a determined attempt to give a formally satisfactory definition. Carnap introduces "intensional isomorphism" as an interpretation of synonymy, defining two expressions as intensionally isomorphic only if they are both logically equivalent as wholes, and have corresponding constituents that are logically equivalent. It turns out, however, that corresponding primitive constituents, such as predicates, for example 'human' and 'rational animal', can be logically equivalent only if the rules of designation where they are introduced show that they mean the same. From our point of view this is obviously unsatisfactory. It is indeed apparent that Carnap is not really concerned, in spite of his claims, with natural language, but with the rather different problems of the relations between complex expressions in formal deductive systems. The point is that the kind of system that the logicians are interested in is too strong for our purpose. We need a much more flexible system for dealing with the complexity and untidiness of natural language, but if possible one which we can describe formally; and the problem is to construct a system that is both flexible, or weak, enough and is still a formal system.

Quine in *Word and Object*[12] has attempted to define synonymy in a way that appears to be more relevant to natural language, by introducing the concept of "stimulus synonymy," or sameness of "stimulus meaning," where stimulus meaning involves both affirmative stimulus meaning and negative stimulus meaning depending on the language-user's reactions to proposed associations of stimuli and verbal responses. Establishing stimulus synonymy for translation between languages involves both careful observation of language-users and analytical hypotheses in which equivalences or correlations between the languages are posited; but, Quine argues, there is always the indeterminacy presented by the fact that different and incompatible sets of correlations are possible, with the consequence that it is very difficult to make sense of the notion of synonymy itself.

This conclusion, however, is not as serious as it appears to be. In one sense it is quite true, but it is a philosophical conclusion, and in practice we do assume that we know what synonymy is, and can set up the correct equivalences, that is, can reasonably say that two words are synonymous. A rather different point is that while Quine correctly bases the attempt to establish synonymy on a careful and scientific investigation of the language-user's behavior, he does not provide the detailed account of a procedure for establishing synonymy quickly and non-contentiously that we require. A further point is that Quine, though he is interested in natural language, appears to be hankering after synonymy in the strong sense in which logicians have tended to interpret it, namely as "total" synonymy; for logicians in general, two words 'A' and 'B' are synonymous if 'A' is always substitutible for 'B' and vice versa. This view of synonymy is apparent, for instance, in the recurring use of "bachelor" and "unmarried man" as an example. Quine indeed admits that words may have different translational synonyms, but appears to treat this as a sort of deviation from the norm, rather than as the norm itself.* The important point is that that view of synonymy depends on the assumption that words have single, fixed meanings. Without this assumption there could be no question of one word always being substitutible for another, and it is this assumption that makes the logicians' treatment of synonymy so unrealistic. It is an empirical fact that words in natural language have different meanings or uses, and that they may sometimes be intersubstitutible, though they are not always intersubstitutible. This means that synonymy is a much weaker relation than the logicians would have it; it has to be treated as a relation between word uses, and not as a relation between words.

The most satisfactory attempt to define synonymy from this point of view has been made by Naess in *Interpretation and Preciseness.*[13] Synonymy as a relation that sometimes, rather than always, holds between words, has been discussed by linguists, and it has been assumed that a substitution test by which words are defined as synonymous in relation to classes of contexts is the best method of establishing synonymy (see Ullmann, *op.cit.).* The linguists have not, however, made any attempt to work out this approach in a rigorous and detailed way. The linguistic philosophers following Wittgenstein have also treated synonymy in this way, since they have been concerned with comparing the ways words are used, and in analysing the similarities and differences between these uses. They have, however, in general assumed that the examples given will be sufficient to make the nature of the relationships between the words concerned plain, and have not discussed these notions of similarity or sameness of use explicitly. (For a typical case see Austin's "A Plea for Excuses."[14])

Naess, on the other hand, is concerned precisely with the detailed problems of constructing procedures that will test synonymy in a context or class of contexts, and of defining synonymy with respect to them. In particular, he elaborates various informant questionnaires for establishing synonymy, including one for substitution. Unfortunately, Naess's questionnaires are far too complex for use in practical lexicography, though they are the kind of thing that would be required, in the last resort, for a really thorough investigation of whether a particular pair or set of expressions were synonymous. The other defect of Naess's approach is that he does not give a general definition of synonymy

* Logicians do not, of course, always stick to total synonymy; they may be prepared to accept that a word 'W' may have uses Wl, W2, W3 etc., to each of which their rules apply; but the complexity that would ensue is not sufficiently considered, and the fact that these are different uses of the same word does not appear in the system in a way that is linguistically satisfactory.

in natural language; each of his procedures defines a particular "questionnaire synonymy," though each of these forms of synonymy is rigorously defined, and has the formal properties like symmetry which the logicians are interested in.

None of these approaches, therefore, is appropriate for our purpose. The logicians' total synonymy does not hold in natural language; in the linguists' use, 'synonymy' and 'substitution test' are ill-defined; Naess's questionnaire synonymies do not give us a general definition of synonymy, and his procedure is too complicated. All the approaches taken together, however, suggest that we ought to be able to give a proper definition of synonymy as a relation between word uses by making use of substitution in some way.

The Definition of Use Synonymy

If we want to say that word uses are synonymous, we cannot do it in the abstract; we have to relate the uses to a context. We cannot, that is, say how a word is being used without reference to a context. To define use synonymy, therefore, we have to substitute in context; by doing so, we get a set of substitutible word uses. In this, we are using the notions of "context" and "use" in the way that linguistic philosophers following Wittgenstein do, but unlike them, are using these notions to give us a definite piece of information, about the synonymy relations between particular words. At the same time, we are pinning down the notion of synonymy by asking whether two words are used synonymously in context, and not, much more vaguely, whether two words are synonymous.

Outline of a Formal System

This is not the place to attempt a full-scale exposition of a formal system on this basis. I shall rather give an outline to indicate the general character of the approach adopted. This may appear evasive, in view of my assertion that a formal system of some kind is required, but the point is that the precise details of a proposed notation are less important than the nature of the interpretation of synonymy, and this can be made clear by giving an outline of the main steps that would underlie a more detailed formal exposition, together with examples. We are, moreover, as noted earlier, concerned with trying to construct a formal system that is flexible enough for natural language, and the kind of system that we find ourselves dealing with in this situation turns out to be very weak in the sense that it constitutes a description rather than a calculus. It is thus perhaps better represented by a series of summary statements than by a mass of equations and symbols.

A formal account of synonymy must, if it is to be of linguistic rather than logical interest, be either a reductionist one in which synonymy is defined in terms of mechanically observable facts about texts, or one in which synonymy is defined in terms of some other linguistic relationship or fact that is taken as primitive. This paper does not offer a reductionist account, but attempts to explain synonymy in terms of a relationship, called "sameness of ploy," between sentences; and the possible logical triviality of the explanation of the one in terms of the other should not be allowed to obscure the fact that this is a legitimate way of explicating the notion of synonymy, and of giving us an interpretation of synonymy that we can use for our practical purpose. The system thus starts with sentences, rather than words or word uses, and can be summarized as follows:

A sentence is a delimited sequence of elements that has a "ploy" (the way it is employed).
Consider a class of sentences with the same ploy;
consider the subclass of this class with the same length (i.e. number of elements);
consider the subclass of this subclass with identical elements in all corresponding positions save one, where the elements differ.
The elements in this position will be said to be "parallel."
A class of elements that are parallel with respect to some position in some class of sentences will be called a "row."

The term 'element' can now be interpreted. A sentence is a sequence of word signs; it is also, because it has a ploy, a sequence of word uses. We can therefore give the following definitions:

A "word-sign" is a delimited sequence of characters.
A "word-use" is an occurrence of a word-sign in a ployed sentence.
A "word" is a class of word-uses with the same word-sign.
A "sentence" is a delimited sequence of word-signs representing word-uses.

Dealing with classes of sentences may be correct, but is not very convenient. It is much more convenient to consider one sentence and replacement in it without change of ploy. Instead, that is, of talking about sentences with the same ploy that differ in one element, we can talk about one sentence and the different elements that may replace one another in it without changing its ploy. We therefore redefine 'row' as follows:

A "row" is a class of word-uses that are mutually replaceable in at least one sentence.

In this formal system, therefore, we have word-uses, and not words, as the primary units. A word-use is defined by synonymous word-uses, that is by word-uses that may replace it in at least one context; and since these word-uses, because they are synonymous, that is mutually replaceable, define each other, we obtain sets of synonymous word-uses, or rows. A word is thus defined by the set of rows in which its uses, that is the set of uses with the relevant word-sign, occur.

An important consequence of this approach is that we can make statements about some other relations between words or word-uses on the basis of our initial statements about these synonymy relations. To start with, if we have defined words as synonyms if they may be substituted for one another, that is, may co-occur in at least one row, we can obviously define words as total synonyms if they can always replace one another, that is always co-occur in rows. This is quite straightforward. We can, however, also define likeness between words in terms of the extent to which their uses are synonymous. Thus, if two words co-occur in a large proportion of their rows, we can say that they are very like; if they co-occur in a small proportion, we can say that they are less like. We can, moreover, make statements about the likeness of two words that have no synonymous uses, in terms of the extent to which they are synonymous with a third common word, and so on, with the likeness diminishing as the number of intermediate words increases. The important point, however, is that we can make these statements about likeness precise; we can measure the likeness between words, and give it a numerical value. This is because we are dealing with numbers of rows. We can say that the likeness between two words is some suitable function of the number of rows in which each occurs and the number of rows in which they co-occur. This can then be modified to deal with the cases where the words do not themselves co-occur.

This development from the initial statements about synonymous uses can be carried further, for example to define unlikeness as least likeness, and so on. We shall not go into this question further here, since it is not immediately relevant, but will only stress the fact that we can build up a complicated picture of the various relations between words, which we can describe as a picture of the semantic structure of the vocabulary, from very simple initial information. We can also obtain further information about various relations between word-uses, rather than words. We shall not, however, consider this point here either, as it is discussed in detail later.

Returning now to our main problem, the rows we obtain by carrying out replacement will be the units for the higher-level classification that gives us our thesaurus groupings; the latter will thus be classes of classes of word-uses. We can say that rows are satisfactory as definitions of word-uses since they are easily handled, concise, precise, and adequate as a means of distinguishing and specifying the various uses of a word. In comparison with other approaches to synonymy, we have on the one hand defined synonymy formally, but in a realistic way as a relation between uses, and on the other, though the method relies on linguistic context as the proper source of information about the way words are used, have devised a procedure in which there is no need to record contextual details explicitly.

Collecting Synonymy Information

The initial data we require in order to construct our thesaurus will thus be sets of synonymous word-uses, with replacement in context as operation for collecting them. To consider the question of collecting our data in more detail: can it really be done? Can this kind of refined analysis of the way words are used be carried out quickly, efficiently, and objectively?

To start with, there is no point in trying to do it, as it were, in the blue; we can use any good existing dictionary like the large *O.E.D.* This is clearly an advantage, as a detailed dictionary of this kind contains a great deal of valuable information, and we can save ourselves a lot of trouble if we can use this information in a straightforward way. If we look at the *O.E.D.* for example, we find that a great many of the entries are virtually rows, and can be "lifted" without modification. This means that row making is quite quick and easy. The *O.E.D.* also gives illustrations of the uses taken from actual texts, and these are ready-made replacement frames.* To give some examples:

"*Act* 1 a) A thing done; a deed, a performance."
Quotations illustrating the use are given:
"As worthy an act as ever he did"; "The prowess and worthy acts of the Ancient Britons"
In both of these examples we can plausibly substitute 'deed' for 'act':
"As worthy a deed as ever he did"; "The prowess and worthy deeds of the Ancient Britons"
"*Act* 4 The process of doing; action, operation."
Quotations given are:
"Wise in conceit, in act a very sot"; "The rising tempest puts in act the soul"; "And hear the flow of soul in act and speech"
In all of these we may substitute 'action' for 'act'. We can also (this is confirmed by checking the entry for 'operation') replace 'act' by 'operation' in the second example, thus obtaining a three-word row 'act action operation' as well as the two-word row 'act action'.
"*Toil* 3 a) Severe labour; hard or continuous work or exertion which taxes the bodily or mental powers."
One quotation is:
"You are many of you accustomed to toil manual; I am accustomed to toil mental."
As the definition suggests, 'labour' can be substituted for 'toil'.
"*Task* 3 A 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'."
One quotation is:
"He had taken upon himself a task beyond the ordinary strength of man."
Here we can substitute 'labour' to get the row 'task labour'.

These examples show how rows can be set up, and how an existing dictionary can be used. The *O.E.D.*

* The formal system requires that a replacement frame must be a sentence (assuming that any stretch of text bounded by full stops — with allowances for abbreviations — is de facto syntactically a sentence). The *O.E.D.* quotations, on the other hand, are frequently not sentences. We can nevertheless use them in practice, as most of the examples could be turned into sentences without any change in their character: thus we can turn 'as worthy an act as ever he did' into 'It was as worthy an act as ever he did'. So long as this could be done in an acceptable way, there is no harm in using the *O.E.D.* examples as they stand, provided that they are full enough to establish a context for the word in question. Using pieces of text that are not sentences is thus simply a matter of practical convenience, and does not affect the formal basis of the system.

definitions are sometimes not very row-like, but they can usually be converted without much difficulty. The entry for 'toil'—'hard or continuous work or exertion which taxes the bodily or mental powers' gives the row 'toil work exertion'. The quotations in the *O.E.D.* are often rather unsatisfactory substitution frames, often because they were chosen for etymological reasons, and they do not allow all the substitutions the definitions suggests. This does not matter, because we are not primarily concerned with the sentences, so one uses them where one can, and if they cannot be used as they stand, they may still be helpful in suggesting other more appropriate sentences for replacement. In practice one does not have to find a context to test each potential row; one's familiarity with the language, and knowledge of the kind of context which would be relevant, is usually sufficient.

The results obtainable can be more fully illustrated by the set of rows for the word 'act', which are part of a larger sample being used for experiments:

act doing
act working performance operation
act achievement
act result outcome consequence
act event
act fact
act thesis dissertation
act statute
act record
act judgement decision verdict
act order command fiat decree
act decree law
act scene
act performance
act pretence sham
act show
act impersonation
action act
operation act performance
performance action act deed operation
performance action act deed
deed act
deed doing act action
deed act action
deed instrument act
proceeding act
proceeding action act
acting act
work act deed
work act

We have constructed rows on this basis without much difficulty, and quite quickly. The method is very simple and does not seem to present any practical problems.* The procedure is of course not mechanized, but it reduces the area of choice open to dictionary-maker to very narrow limits. The only way of extracting linguistic information without any intervening human judgment is by the mechanical scanning of text, but this is well-known to be exceedingly inefficient as a method of obtaining semantic information, and it is in any case difficult to see how it could produce rows.

The method can still be criticized in two ways. It may be maintained, firstly, that no two words are ever replaceable without change of ploy in any context, and secondly, that two words are always replaceable without change of ploy in some context. In answer we can say, firstly, that we are dealing with uses, and not words. The overtones of two words, representing their whole ranges of uses, will nearly always be different, but in a particular context their uses may, for all practical purposes, be indistinguishable. This is not very satisfactory, but can be supported by the empirical argument that we (ordinary language-users, that is) do say that words mean the same in particular contexts, and substitute them. We can say, secondly, that while one can always construct a context in which any two words are replaceable without change of ploy (a great many words can be unhelpfully replaced by 'thing'), one has to work quite hard at constructing a context that is both far-fetched and plausible; and the practical dictionary-maker is concerned with the ways in which words are ordinarily used, and not with playing games with language. The real point is that though we have to depend on the language-user somewhere, in this approach the subjective element is restricted as much as possible; the dictionary maker has only to decide whether 'A' can replace 'B' in context x. This is not strictly objective, but in thus saying that the method is not wholly objective, we are not making a very damaging admission. In contrasting "objective" and "subjective" in language analysis we are in theory contrasting methods that can be carried out automatically and methods that rely on a human language-user, or informant, or dictionary-maker, at some stage. But this is a somewhat irrelevant distinction, since no one has yet succeeded in making a *dictionary,* that is a dictionary defining the meanings of words, without any human intervention (say by scanning text mechanically, *and* sorting and evaluating the results obtained mechanically). In practice one is concerned with what maybe called "intersubjective validity"; does the human being involved produce results that are generally acceptable? This is, I claim, best achieved if we pin him down to a particular decision about the particular use of a particular word, instead of asking him for the possible uses of a word.

Testing Replacement in Context

The criticisms just discussed suggested a small-scale experiment to test the replacement criterion. This was carried out on Richards' and Gibson's *English through Pictures,*[15] which is a teach-yourself book containing simple sentences with an explanatory diagrammatic picture for each one. As every sentence is tied to a picture, it can be unambiguously interpreted, and as

* The examples just given are rows for nouns, but rows for other parts of speech have been and can be constructed. An important feature of this method of indicating the meanings of words is indeed that it can be applied to any kind or class of word; thus we may have the rows 'to towards', 'each every'.

the sense of the sentence is pinned down by the picture in this way, one can really decide whether a word in it can be replaced by another or not. Rows were obtained by carrying out replacement, where possible, for every position in every sentence in the book, for example as follows:

She put the hat on the table
She placed the hat on the table

The character of the rows obtained can be illustrated by an example:

bit piece
bit lump
crush mash
ready prepared
sort kind
dry wipe
round circular
round globular
push jog
fall tumble
fall drop
good thorough
good efficient
good comfortable
good pleasant
good satisfactory
good first-class
good nice

The experiment was in fact not very satisfactory. The sentences are often so simple, for example, 'This is a hat,' that there is no opportunity for replacement. Many of the words, such as 'apple', are names of physical objects, and these, unlike 'action', are the least replaceable words in the language. There are also, in contrast, a small number of words, like 'do', that are used in an unnaturally large number of ways, as in Basic English. (This can only happen where there are pictures to give a precise interpretation.) We therefore obtained a very small number of rows for many words, and a very large number for a few words, and this gave a very unbalanced sample. The experiment did, however, show that replacement can be carried out in a quite straightforward way without doubt or difficulty.

The procedure for carrying out semantic analysis just described gives us, as our basic semantic material, sets of synonymous word-uses. In each set, or row, a use of the words concerned is defined. Now it is clear that analysis on this level of detail will give a very large number of rows, and that some sort of organization and classification would be required, even if we were not trying to construct a thesaurus. We are, however, specifically concerned with constructing a classification of the fundamental kind represented by a thesaurus, and the question we now have to consider is how we obtain such a classification.*

A Possible Approach to Classification

One approach is to apply the Theory of Clumps.[16]† In clumping, objects are classified on the basis of their properties, using an initial data array of the following form:

$$\begin{array}{cc|ccccc}
 & & \multicolumn{5}{l}{\text{Properties}} \\
 & & P_1 & P_2 & \ldots & & P_n \\
\hline
\text{O} & O_1 & 1 & 1 & 0 & 0 & 0 \\
\text{b} & & & & & & \\
\text{j} & O_2 & 1 & 0 & 1 & 1 & 0 \\
\text{e} & & & & & & . \\
\text{c} & . & 0 & 0 & 1 & 1 & 1 \\
\text{t} & & & & & & . \\
\text{s} & O_n & 0 & 1 & 0 & 0 & 0
\end{array}$$

where O_1 has P_1, P_2, O_2 has P_1, P_3 and so on. Using some similarity or association coefficient, we compute the similarity between a pair of objects on the basis of their common properties. In the semantic case the rows are clearly the objects. But what are the properties? The only possible properties which a row can have are the word-signs which occur in it. For example, consider two rows A B C and A E F. A in each row is the same sign; and A in each row represents a use of the same word, because we defined a word as the class of uses with the same sign. The trouble is that this is a formal definition of a word. The fact that the sign occurs in different rows means that it represents different word-uses, and the fact that these uses have the same sign means only that there is the formal relation between them of having the same sign. What do we know about the semantic relation between two uses represented by the same sign that would

* It must, however, be emphasized that the method of analysis we have described can be used without any reference to further classification to give a thesaurus. We can, for example, if we wish to construct an alphabetical dictionary, set up our rows, and then, given our words in alphabetical order, distribute the rows so that each row is listed under all the words that occur in it. This approach to semantic analysis is thus quite general, and need not be geared to the construction of a thesaurus. Given that very refined dictionary-making is required for high quality machine translation, the procedure described has the advantage of being simple and rapid, and of distinguishing and defining the uses of words in a very efficient way.

† The Theory of Clumps has been applied primarily because classification programs based on it are available in Cambridge. It might turn out that this approach is not the most suitable for the semantic material with which we are concerned, but as we do not know what a more appropriate procedure should be like, we can only try existing procedures and see how they work out. The Theory of Clumps is in any case intended to be a general theory of classification, which may be applied in quite different fields, so it can reasonably be applied in this field. A further point is that the procedure is both simpler and more applicable to larger quantities of data than others that are being developed.

make it possible to regard the occurrence of a sign in different rows as semantically significant? We call the uses represented by the same sign the uses of a word; what does this imply? If word-uses are our primary units, how can we connect them other than by their signs?

The Economy Hypothesis

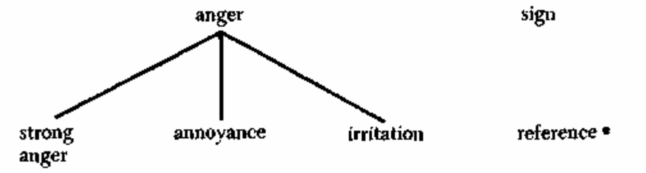
To answer the question just posed, we have to examine the nature of language in general. We can say, very crudely, that a language (strictly, a vocabulary) is a set of signs that represent a set of extra-linguistic references or situations, using 'reference' in the widest sense. Now consider a language with one sign per reference (or a number of references that are regarded as identical for practical purposes). We might, for example, have a language that used the sign 'shule' for the reference "shoe," the sign 'sindle' for the reference "sandal," and the sign 'griss' for the reference "grass."* The International Code of Signals is essentially a language of this kind. In the Code each sign is unambiguous, that is, has a unique reference (or type of reference). The Code is, however, a very limited language. It deals with a very limited number of highly stereotyped references and situations. If we had one sign per reference, and had to deal with the vast number and variety of references with which an effective natural language must be concerned, we would have far too many signs; the language would not, humanly speaking, be manageable. Some kind of sign economy would be required.

We can now consider how this economy might be obtained. Consider a language in which a sign stands for a set of very different references. We might, for instance, using the previous example, use the one sign 'shule' for the two quite different references "shoe" and "grass," so as to eliminate the sign 'griss'. There will be no (or virtually no) ambiguity, because the surrounding context will distinguish the relevant use of the sign; it would be as if the language consisted of systematic homonyms. This device would effect the necessary economy, but a language of this kind would still not be very manageable from the language-user's point of view. There would be nothing characteristic or coherent, and therefore memorable, about the meaning of the sign. Now consider an alternative language in which a sign stands for a set of similar references. Thus, we might use the sign 'shule' for the references "shoe" and "sandal," and perhaps also for "brogue" and "boot." This would be manageable, as there would be something consistent or coherent about the way a sign is used, about its meaning or interpretation. This is, I maintain, what we mean when we talk about a word and its range of uses. It may not be that any two uses are very close, but it will be true that each use will be close to one or more of the others; there will be, metaphorically speaking, a continuous series of uses. Particular uses will again be distinguished by context. They can also, as we have suggested, be distinguished by their synonyms.

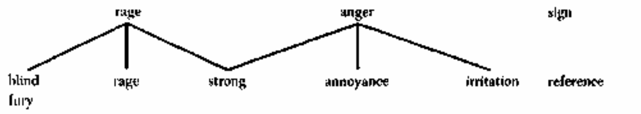
If we adopt the third approach we can effect an economy in the number of signs required without putting a limit on the number of situations with which the language can deal, and we can obtain this economy in a very efficient way. What we have is a hypothesis, which we shall call the Economy Hypothesis, to the effect that as we have to use one sign for several references, we use a sign for similar references. We are, however, still left with the question: why are there synonyms, that is, synonymous uses, in language? If we can distinguish uses by context, why should we be able, as in practice we are able, to distinguish them by synonyms as well? Synonyms are apparently redundant and unnecessary. If so, why do we have them?

The Synonymy Hypothesis

Consider the model just described. When we group together a set of references or situations to be represented by one sign, we are emphasizing one characteristic or common feature of the references concerned. We can illustrate this as follows:



In fact, these references or situations have different aspects, that is, can be looked at in different ways. (Putting it crudely, nearly everything can be looked at from more than one point of view.) If these references only occur in one sign group, therefore, they are, in some sense, inadequately represented in the language. If they are to be properly represented, we should pick up their other aspects; the references, that is, should occur in other groups represented by other signs, where other features of the references concerned are emphasized. This can be illustrated as follows:



This means that for the reference "strong anger," which will be a particular reference in a particular context or

* The references cannot strictly be represented by words other than 'shule', 'sindle', and 'griss'; we are using "shoe," "sandal," and "grass" simply as labels in the absence of the actual extra-linguistic references.

* The references cannot strictly be represented by words other than 'anger': we are using 'annoyance', etc., simply as labels in the absence of the actual extra-linguistic references for them.

contexts, two signs will be equally appropriate; either 'rage' or 'anger' will do. 'Rage' and 'anger', that is, will be synonymous in this particular case. The ranges of references represented by 'rage' and 'anger' respectively, however, will be different.

The argument, then, is that when we assign individual references to groups of similar references, to be represented by a particular sign, we find that we wish to assign a particular reference equally to several groups because it is similar to references in different groups, in different ways, and assigning it to different groups means that we have several different signs for it. The groups themselves are distinct, so that there is a genuine difference between the signs, with respect to the groups, but there is no difference between the signs with respect to any single common member of the groups. When we are concerned with that particular reference, we can use any of the relevant signs indifferently. At the same time, most references will not be members of identical sets of groups, and so will not be represented by identical sets of signs. We thus distinguish a particular reference from others by its being represented by a particular set of signs, and at the same time define it by this set of signs. These signs, when they appear in ployed sentences, represent the uses of words, so that the fact that a particular set of signs, or word-uses represented by signs, can indicate a particular reference, means that we have a set of synonymous word-uses.

This argument thus suggests that synonymy is a fundamental feature of language. If we do not have any synonyms, it means that the grouping of references under signs is incomplete. We thus have another hypothesis, which I shall call the Synonymy Hypothesis, that says that different words will have uses that stand for the same references, so that their signs are equally appropriate where these references are concerned, and that explains why we can hope to find rows and get a useful semantic classification out of them. This is because synonymy relations between words reflect the way we look at extra-linguistic references.

To revert to the earlier problem of classification. The Economy Hypothesis justifies the belief that there is a semantic relation between word-uses with the same sign, and therefore between the rows in which they occur. This is a general remark, that is, it is in general true that two word-uses with the same sign will be semantically closer than two uses with different signs. We cannot measure the closeness or likeness precisely, and it may not be true in particular cases. However, if it is true in general, that is, for any two uses with the same sign considered in relation to the language as a whole, we can measure the similarity or "overlap" between rows in a precise way. We can justify the assertion that rows with a common sign have something semantic in common, and therefore that the greater the number of signs in common, the closer the relation between the rows concerned.

Classification Experiments so far Carried Out

For experimental purposes, a row sample based on the *O.E.D.* was prepared. The chief difficulty is obtaining a sample which is both small enough for computer handling and reasonably representative. To see how rows are related to one another one has to have a number of rows for some words—if possible all the rows for some of them,—and also rows for a number of words—if possible for some words that define each other. Experiments so far have dealt with 500 rows, but 2000 have been prepared. For the initial sample of 500 a small number of words that we have called "starting words,"* with varying ranges of uses, but with some uses in common with some of the others, was selected. All the rows for each of these words were then worked out. This meant that in the sample as a whole there were some words for which all the uses were given, some for which some uses were given, and some for which only one or two uses were given. There were some starting words that co-occurred several times, and other words that occurred only with a particular starting word. The starting words were: 'act, action, activity, business, operation, performance, task, labour, toil, deed, effort, creation, product, production, function, conduct, proceeding, acting, work, working'. Their sets of rows ranged from 19 for 'acting' through 48 for 'business' and 49 for 'operation' to 90 for 'work'. 325 other words were involved; 200 of these only occurred once, 67 twice, 19 three times.

These figures show that the sample was not very satisfactory. There were far too many "once words" compared with those that occurred more often. This is clearly unsatisfactory, since the words concerned do not in fact have only one use. An attempt to remedy this was made by taking all the words that co-occurred with 'work' and setting up all the rows for them. This gave a further 1500 rows.

We have seen that the occurrence of word-signs is a significant property for computing the similarity of two rows. The next problem is to find a suitable similarity or resemblance coefficient. For the first experiments one that had already been used for other experiments in grouping was taken over. In terms of objects and properties, this is defined as follows:

$$S_{o_1,o_2} = \frac{\text{Number of properties in common}}{\text{Total number of different properties}}$$

In this case we have rows as objects and signs as properties. Thus if we have the two rows 'action act' and 'deed act', for example, their similarity is 1/3, and if we have 'performance action act deed' and 'operation act performance' we get 2/5. The initial data array of the form given earlier is converted into a similarity matrix for pairs of objects, in this case pairs of rows,

* We have used this rather horrible phrase, rather than, say, 'key-words', as we do not wish to suggest that these words have any special semantic character. They are simply the words that were completely analysed for the purposes of the experiment.

and the group-finding operations are carried out on this.

Given our similarity information, we have to have a definition of group, and a procedure for group-finding. Roughly, we want to define a group as a set of objects that are more like one another than they are like non-members. Very different definitions will meet this specification. The particular one adopted is taken from the Theory of Clumps, where it has been used in a number of fields. The definition is as follows:

A subset is a group, or "clump," if each member has a greater total of similarities to the other members than to non-members, and vice-versa for non-members. In the clump-finding procedure the total set is partitioned and iteratively scanned, elements being redistributed after each scan until a satisfactory similarity balance is achieved.

The first clumping experiments were carried out on a sample of 180 rows. These were satisfactory as far as they went, but the sample was too small for informative results. The next tests were carried out on the 500-row sample. The first runs of the program produced quite a lot of clumps, but they were unsatisfactory in two respects:

1. Many of them were too big; they were aggregates of what one would have hoped would be smaller clumps. (Given the data, there is something wrong with a clump containing 249 elements).
2. The smaller individual clumps, and the subsets of the larger ones, both tended to be simply the sets of rows for a particular starting word. 'Production' and 'work', for example, generated clumps, and one aggregate consisted of nearly all the rows for each of 'act, action, activity, operation, performance, deed, proceeding, acting, working'.

The trouble with clumps that are centered on particular words is that, although the uses of a word have some relation to one another, the relation between every pair is not necessarily very close. In particular, it is not necessarily as close as the relation between one of them and another row that does not contain the word concerned but does contain other common elements. It was also the case that in many of these clumps some of the rows containing the focal word did not occur. Thus, the row 'production work' did not occur in the clump centered on 'production', although one would have said that it should be there. This turned out to be because 'production' came in 43 rows in the sample, whereas 'work' came in 90. This meant that the row 'production work' had a greater total of connections to rows containing 'work' than to those containing 'production', that is, had a greater total of connections outside the 'production' clump than inside it. This sort of thing occurred in more subtle forms elsewhere. Groups of rows that one would have said should have come together failed to do so, because the total of the external connections of the members was greater than that of the internal ones. Thus, the

staging production
acting staging production
staging production performance
production performance
acting production performance
staging performance
acting staging
acting performance

failed to come as a separate clump because the "pull" of outside rows containing 'production', 'performance', or 'acting' was greater than the internal coherence of the clump.

Now it is clear that the simple number of uses of a word should not be allowed to affect grouping in this way. The similarity definition was therefore altered so that the similarity between two rows is dependent on the frequency of the words in the rows: similarity in a frequently-occurring word counts for less than similarity in an infrequently-occurring word. Thus if the word 'work' is common to two rows it contributes only l/90th, not 1, to the similarity; but if the word is 'organization', it will contribute 1/2 instead of 1.*

Further experiments were carried out with this revised definition. In contrast to the earlier experiments, the results were satisfactory in that the clumps were not aggregates or centered on starting words, and they were also satisfactory in that there were some plausible clumps, on an intuitive evaluation. The set of rows containing 'acting staging production performance' listed above appeared, and the following rows also came out as a clump:

action activity briskness liveliness animation
activity animation
activity liveliness animation
activity animation movement
activity briskness quickness liveliness speed
activity motion movement
activity movement business
activity movement
business briskness liveliness

In both cases one would say that these are thesaurus-type conceptual groupings; they can be given headings like "Staging" or "Animation." Thus, though the experiments carried out so far have not been very extensive, the results obtained do suggest that we can derive thesaurus groupings from our initial data by a purely automatic procedure. This last is most important, not merely because it enormously reduces the amount of effort involved in constructing a thesaurus, but because it means that the groupings are objective. We cannot construct a thesaurus by wholly objective, i.e., automatic, means; we cannot abolish the subjective element in lexicography entirely; we have to depend on the language-user's judgment somewhere. But in setting up rows, he exercises his judgment within very rows:

* To put it more precisely: where previously a word contributed 1 to the various counts used in computing a similarity, it now contributes 1/N, where N is the total number of its occurrences.

restrictive limits. He has only to decide whether two words are mutually replaceable without change of ploy in a single context. This leaves considerable scope for thought to the dictionary-maker, but he is not being asked merely for a judgment of synonymy; he is being asked to answer a much more precise question. This attempt to minimize the subjective element would, however, be wasted if the subsequent grouping were done intuitively. An automatic grouping procedure is theoretically as well as practically desirable. In saying that the clumps illustrated above are thesaurus-type conceptual groupings, we are making an intuitive judgment, based on a comparison between the clumps and the kind of head in Roget's *Thesaurus* which we originally took as our exemplar. This is to some extent a sufficient reason for saying that our experimental results are satisfactory, but we should perhaps look at this question of conceptual groupings a little more closely. We have assumed that we know what we mean when we say that a thesaurus head in, say, Roget's *Thesaurus,* is a conceptual grouping, but we should inspect this assumption.

The notion of "conceptual grouping" in itself is very vague. As we saw earlier, we could treat Roget's heads either as sets of words that express the same concept, or as words that are synonymous. We were thus treating one kind of Roget head, the synonym group, as typical. There are, however, other heads in Roget's *Thesaurus,* like 267 NAVIGATION or 191 RECEPTACLE. The former contains words for anything to do with navigation, for example 'oar' and 'mariner', and the latter words for any kind of receptacle, on a very wide interpretation of 'receptacle', such as 'oriel' and 'commode'. In some sense these are conceptual groupings, in the way in which closely related headings in a hierarchical classification like the U.D.C. could be said to form a conceptual grouping, but they are rather different from heads like 24 DISAGREEMENT which consists almost entirely of synonyms and near-synonyms like 'disagreement', 'disunion', 'discrepancy', 'divergence' and so on. It can reasonably be said that words like 'oar' and 'canvas' do not express the idea of navigation, or 'closet' and 'nook' the idea of receptacle, in any very precise sense; 'discrepancy' and 'divergence' on the other hand do express the idea of disagreement.

The real difficulty lies in saying that a set of words form a conceptual grouping if they express a particular idea. This is too vague to be useful. It raises too many problems about what it is for a word to express an idea. This does not, however, mean that we cannot give the notion of conceptual grouping a more precise interpretation. If we say that two words can be used in the same way in one or more contexts, that is, are synonymous, we can say that they must express the same idea, without our having to investigate or specify how they express this idea, or, more importantly, what this idea is. If we have a set of words that can be used in the same or similar ways, where sameness and similarity are defined in the way we have described in terms of occurrence in the same row and in overlapping rows, we can say that we have a set of words that express the same general idea. That is to say, we are defining a conceptual grouping as a collection of synonyms and near-synonyms, and not, for example, as a collection of words that stand for a particular sort of physical object. This, then, makes clear both what is meant by the description of one kind of thesaurus head as conceptual groupings, and by the assertion that clumps of overlapping rows represent conceptual groupings: a conceptual grouping is a set of words that express the same idea; a collection of synonyms and near-synonyms must necessarily express the same idea; and as clumps or rows contain synonymous and similar (or near-synonymous) word uses, such clumps must be conceptual groupings.

Reverting to practical questions, the real difficulty in the actual experiments is evaluating the output. One has an intuitive idea of what one wants, namely clumps of the kind just discussed. But this intuitive idea is a general idea, and the problem is to give a detailed estimate of what is right or wrong about a particular clump, not merely in itself, but against the background of the data as a whole. One has to decide both whether there are rows in the clump that should not be there, and rows outside it that should be there, and this is very difficult with such heavily overlapping material. Clumps which contain rows without much overlap do not present many problems. If there is too little overlap, the clump should probably not be a clump, but if there is a lot of overlap, the difficulty comes in keeping track of all the overlaps and sorting out the relations between the rows concerned. We must, moreover, when we are classifying large quantities, or all, of our material, evaluate the classification as a whole as well as the individual clumps. That is, for example, we must decide whether the total number of clumps obtained is correct, given the number of rows. Intuitive evaluation of either particular clumps or the set of clumps is clearly not very satisfactory. Even if what we get looks all right, the real test is whether our thesaurus dictionary works for machine translation. We might have a thesaurus that appeared to be a wholly satisfactory improved version of Roget's and yet turned out to be unsuitable for machine translation simply because this kind of thesaurus is not the right kind for this purpose. The trouble, however, about trying to test our thesaurus in this way is that this involves so many other problems, like choosing the correct alternatives from sets of possible parsings, for which there is no immediately obvious solution, that there is some excuse for just looking at what we get. The current state of machine translation research is such that we cannot hope to test any particular solution to a particular problem within the framework of a general procedure, simply because no such procedure exists. In this situadon, the best we can do is look at our classification output in the context of our original data, and compare it with existing classifications like Roget's, on the assumption that we do want this kind of thesaurus. We cannot, given that we are using different material, and a different procedure, make a detailed comparison with Roget's *Thesaurus.* We cannot expect to get exactly the same heads, but we can usefully compare the general character of our results with the kind of Roget head that we took as our guide. We may also be able to test our output in some kind of thesaurus intersection procedure, though this could only be done in a very crude way, in the absence of the larger translation procedure of which such an intersection procedure was intended to form a part.

Measuring Semantic Distance

The starting point for the work described above was the assertion that a thesaurus-type semantic classification would be required, in machine translation, to resolve semantic ambiguity. The question we have still to consider is whether, given a much better thesaurus than those currently available, a thesaurus intersection procedure will work. It may indeed be that repetition of some kind resolves ambiguity, but it does not follow that the relevant uses of the words concerned are specified by thesaurus heads. Why do we think that this is the correct model of language?

Given that there is some kind of semantic coherence about continuous discourse (to put the point as vaguely as possible), we can say the following: if discourse has some semantic coherence, it must be because the relevant uses of the words in the text are semantically nearer to one another than the non-relevant ones. We can say, that is, that the semantic distance between the uses concerned is less than that between the other uses of the words in the text. This is a very vague remark; we have to give 'semantic distance' some kind of interpretation before it is at all useful. I want to suggest that we can use rows to make the whole thing more precise. Suppose that we say that two rows with a word or words in common are one step apart, and that two rows that are each one step from a common row are two steps apart, and so on. We can then give a very precise measure of the semantic distance between the uses of two words, as represented by two rows, by counting the steps between them. This may not be the only possible interpretation of 'semantic distance', but it is a measure of semantic distance in some sense, and any measure is better than no measure at all.

We can now see how this works out for text, taking sentences as units within which this procedure for measuring semantic distance is to be carried out. Suppose we consider, as the simplest case, a two-word sentence 'AB' (disregarding problems about parts of speech). In this procedure we consider the rows in which 'A' and 'B' occur. If they co-occur in a row, this is shortest possible distance, as there are no steps from one to the other; we can illustrate this (rather trivially) as follows:

A B

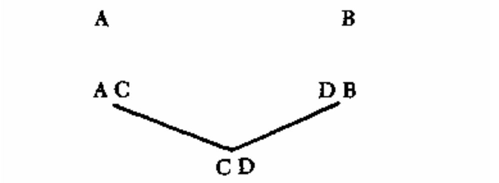
AB

If A and B co-occur with a third common word C, we get a one-step link:

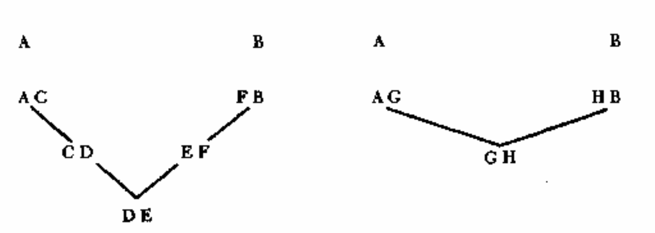
A B

AC———CB

If A and B co-occur with C and D respectively, and C and D co-occur, we get a two-step link:



In each case, we are concerned with the distance between particular rows defining particular uses of the words A and B. The argument is that if we have alternative "routes" from one text word to another, through different series of rows, those rows for the text words that form the end points of the shortest route, and therefore specify the least distance between the text words, specify the correct uses of the text words. Thus suppose we have our sentence AB, and have two routes from A to B, as follows:



There are 4 steps between A and B in the first case, and only 2 in the second, so that the semantic distance between A and B is less in the second case. We can therefore, given our text words A and B, and the information that each can be used in the two ways represented by the rows AC and AG, and FB and HB respectively, say that the correct uses of A and B are those specified by AG and HB because they are nearer than AC and FB.

To test this hypothesis, we have to take words in sentences, examine alternative routes between them, and see whether the uses giving the shortest routes are the correct ones. A number of hand experiments on these lines have been carried out. These were not very efficient, since finding the shortest route between two words depended on knowledge of the row sample, but it was thought that the "route-finding" procedure should be tried on a small scale before extensive computer experiments were put in hand. For the experiments, sentences using words in the 2000-row sample were constructed. These were quite straightforward—there were too many rows involved for there to be much danger of fixing things so that they would work. As the sentences had to be realistic, other words were included. This meant that the procedure could not be carried out for all the words in the sentence, but this did not matter as the point of the experiment was to see whether the correct uses of any words could be selected.*

On the whole the experiments were quite successful. To give some examples:

The *calculations* his *work* involved were enormous

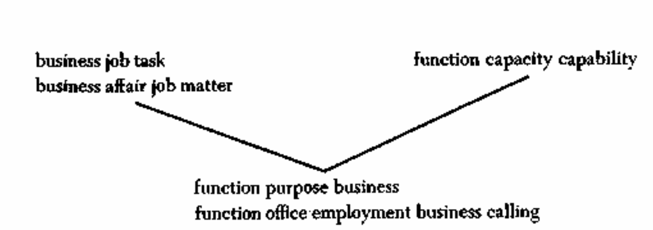
work calculation
work calculation sum
work working-out calculation

Here the two words co-occurred. The sense of 'calculation' selected is quite correct: one could say “The sums his work involved were enormous.” The use of 'work' specified is, however less plausible, though it is more obviously in the right area than 'work' meaning, for example, "fortification."

The *mine* was in full *production*
work working mine ————— work production

Here there is a common third word, 'work', so there is a one-step connection. The sense of 'mine' specified is quite correct, as opposed to, say, that defined by 'landmine', and so is that of 'production', as opposed to, say, the use defined by 'performance staging'.

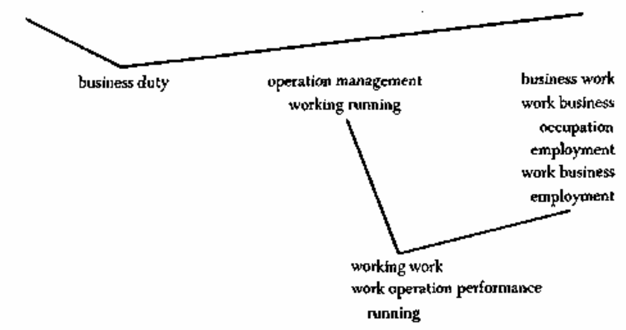
The *job* was beyond his *capacity*


business job task
business affair job matter
function capacity capability
function purpose business
function office employment business calling

Here there is a two-step route via 'business' and 'function'. 'Job' is indeed being used in the sense of 'task', and 'capacity' in the sense of 'capability'.

There were also more elaborate sentences, for example one with three-way links as follows:

His *duty* was the daily *management* of the *business*



'Business' and 'duty' co-occur, while there is a two-step route between 'business' and 'management' via 'working' and 'work'. The senses of 'duty' and 'management' are correct. One can substitute 'business' for 'duty' and 'running' for 'management', and the sense of 'business' defined by 'work' is nearer the mark than that defined by, for example, 'animation'.

The following is one that did not work so well:

The *ideas* in his recent *work* are remarkable

'Idea' is defined by 'notion', and 'work' by 'invention', and 'notion' and 'invention' co-occur. The sense of 'idea' is correct (there were other defining words like 'theory' as well), but 'work' does not mean 'invention'. It can, however, be said that 'work' means 'invention', that is, that we are in the conceptual area labelled “research” or “investigation,” rather than “mine” or “needlework.” From this point we can indeed draw a general conclusion.

The practical difficulty about the model of semantic distance we have just considered is that whether we get the correct result or not in any actual example depends on whether the dictionary maker has made all and only the correct rows, and as we cannot be sure of this, the model is in the absolute sense untestable. This would not, however, really matter if we were careful in our dictionary-making and did a large enough number of experiments. A much more serious point is that the model itself has two defects. It is far too complicated; surely we do not go through all these detailed calculations every time we understand a text. It is also the case that the selection of the correct use is too much of a hit-or-miss affair; it is conceivable that, given two routes between A and B of 27 and 28 steps, say, that we would intuitively say that the second route, though longer, actually specified the correct uses on any independent interpretation of the text (for example, by taking extra-linguistic references into account). Some simpler model is surely required.

We defined semantic distance in terms of routes through overlapping rows. We would say that the rows A C and B D are very close if they are linked through C D. We would, however, also say that two rows that occur in the same group or clump of rows are close to one another, simply on the basis of our

* More properly, this would not matter if the experiments failed; but it would matter, though not very much, if the experiments were successful, for the following reason: suppose that we are considering only two words from a sentence, and that the one selects the correct use of the other. It could happen, if we considered all the words in the sentence, that other routes for these words selected other uses of them. For example, in the sentence, ABC, the route to B selected one use of A, and the route to C selected another. This, however, brings up the question of whether we carry out our route-finding procedure within a sentence on the basis of some pattern or other, and as finding the correct pattern or set of patterns is a major problem in itself, there is a great deal to be said for investigating the route-finding idea itself first, though in an oversimplified and incomplete form.

requirement that a clump should consist of similar rows, where similarity is defined in terms of overlap between rows. We might indeed find that A C and B D occur in the same clump, together with C D, which is similar to both and so brings them into the same clump. Suppose now, therefore, that we have our two text words A and B with their respective sets of rows, and that with the route-finding procedure we find that there is one 3-step and one 19-step connection between them. If we also have a set of groups available, we may well find that the two uses of A and B selected by the first route are specified by rows that come in the same clump, while the other uses are defined by rows in different clumps. That is to say, if we replace our words A and B by the two sets of clumps in which their rows occur, we will find that one clump occurs in both sets, and that the rows defining the uses of A and B selected by the shorter route both occur in this clump, whereas the uses selected by the longer route are defined by rows in different clumps.* In doing this, we have replaced the sets of rows for each word by the sets of clumps which these rows occur in, and have then carried out a set intersection procedure on the latter to find a common clump; this has given us the same result as with route-finding procedure, but we have obtained it with very much less effort.

The substitution of a clump-intersection procedure for the route-finding procedure thus deals with our first problem; we have found a model of semantic distance which is simpler than that on which the route-finding procedure is based. This intersection procedure should also deal with the problem of "near-misses" in specifying the correct use. This is brought out by the last example, showing the case where the route-finding procedure did not work properly. In this example, we obtained the specification of 'work' as 'invention', which was not quite correct, but which we could say was in the correct area of meaning, since we are concerned with ^{c}work' in the sense of 'research' rather than 'work' in the sense of 'needlework'. Now though we may doubt whether the nearest uses of A and B will always be the correct uses of A and B, it is extremely probable that the correct uses will be nearer than the incorrect ones. That is to say, if we have three uses of A that are 7, 8 and 19 steps from B, and if the first use is not correct, the second as opposed to the third will be. The trouble with the route-finding procedure is that it will only give us the first use, though this may be in the right area of meaning and not wholly wrong.

Suppose now that we have clumps of rows, and carry out our intersection procedure. If the first use of A is in the right area of meaning, and the second is the correct use, the rows representing them may well fall in the same clump, so that the clump-intersection procedure would pick out both these uses, the correct one as well as the nearly correct one, and would exclude the third wrong one. The intersection procedure would thus again give us a better result than the route-finding procedure, essentially by being less refined, so that we are more likely to obtain the right row along with others in the right area of meaning. It would, of course, in this case give us more than one row, though this would not always happen, but as the route-finding procedure can also give us several rows for one word which are equidistant from another, as is shown by the examples, this is not a defect of the intersection procedure alone. The number of rows obtained is to some extent a function of the degree of refinement of the row classification, but we could easily have several rows for a word in one clump, with quite a crude classification. Perhaps the best way of dealing with this result is to regard all the rows within a clump as one row. There will after all be no discrimination in terms of the clump classification. This would correspond to the situation where the route-finding procedure selects several close rows, but would eliminate rows that are selected as equidistant but which do not come in the appropriate clump.

We have thus replaced the complicated route-finding procedure by a much simpler and more reliable clump-intersection one. Instead of looking for the links between individual rows, we operate with groups of rows and look for the links between them. We look not at the way words occur in rows, but at the way rows occur in clumps. We have said that the rows in a clump come in the same area of meaning, and we saw earlier that we can say that a group of overlapping rows represents a conceptual grouping, so that we are looking in our intersection procedure for conceptual repetition. We have also argued that these groups of rows are thesaurus heads of the kind we required, so that what we have is a head-set intersection procedure like the one with which we were originally concerned. What the foregoing argument gives us, therefore, is some justification for thinking that a thesaurus-head intersection procedure will resolve ambiguity.

One point about this argument is particularly important: we can see intuitively that “concepts” recur in discourse. In “He went to the bank to cash a cheque for five pounds” we would say, putting it as informally as possible, that the idea of money keeps coming through. But when we interpret 'concept' as “thesaurus head,” this as it were makes a concept a very definite unit, and when we interpret conceptual repetition in terms of recurring thesaurus heads, we are making the vague notion of conceptual repetition very definite too. If we regard a thesaurus head as a set of words that all come under a particular heading, and set up a thesaurus model on this interpretation of a head, with a list of headings, therefore, we are making a number of quite strong assumptions about what a concept is and which concepts there are, and about the nature of discourse, and it can be argued that this is undesirable.

* On some definitions of clump this might be provably so, but the clump definition used was adopted without this in mind.

In contrast, our model of semantic distance, as represented by the route-finding procedure, follows directly from the very simple method of describing the uses of words by rows, and does not essentially depend on the repetition of notions or concepts. The use of an intersection procedure is then only a simplification of the initial model, which makes use of the groups of rows that exist in the set of rows for a vocabulary and that are specified without any reference to concepts. We are thus starting with a procedure to resolve ambiguity by measuring semantic distance that does not depend on any assumption about any *a priori* semantic entities of the kind represented by headings or conceptual classifiers. At the same time, we can see how a thesaurus-type model grows naturally out of the initial one.

To put this point in another way: if we try for head intersections, the procedure may or may not work, though if it does, we can see why, but there is nothing in the heads themselves to suggest why they ought to repeat. Our model, if it works, gives us a reason for thinking that the head-intersection model will work too, that is, it tells us why it should work. We are thus presenting a non-repetitive model, and then deriving a repetitive model from it, and this means that the criticisms that can be brought against the repetitive model can be avoided, just because it is derived from the non-repetitive one. This is not to say that there are no assumptions behind our model, but only that they are less offensive, because less sweeping, than those on which the repetitive one is based.*

* The work described in this paper is more fully developed in the author's Cambridge University doctoral thesis.[18]

Received August 11, 1964

References

1. Masterman, M., “The Potentialities of a Mechanical Thesaurus,” read at the International Conference on Machine Translation, M.I.T., 1956, abstracted in *Mechanical Translation,* Vol. 3, No. 2, 1956.
   Masterman, M., “The Thesaurus in Syntax and Semantics,” *Mechanical Translation,* Vol. 4, Nos. 1/2, 1957.
   Masterman, M., “Translation,” *Proceedings of the Aristotelian Society,* Supplementary Volume, 1961.
   Masterman, M. and Needham, R. M., “Specification and Sample Operations of a Model Thesaurus,” read at the National Physical Laboratory, 1960, mimeo, available from C.L.R.U.
   Parker-Rhodes, A. F., “Some Recent Work on Thesauric and Interlingual Methods in Machine Translation,” International Conference on a Common Language for Machine Literature Searching and Translation, Cleveland, Ohio, 1959.
   C.L.R.U., “Essays on and in Machine Translation,” 1959, mimeo, available from C.L.R.U.
2. Roget, P. M., *Thesaurus of English Words and Phrases,* Penguin Books, London, 1953.
3. Masterman, M., “Semantic Message Detection for Machine Translation, using an Interlingua,” *Proceedings of the 1961 International Conference on Machine Translation of Languages and Applied Language Analysis,* Her Majesty's Stationery Office, London, 1962.
   Masterman, M., “The Semantic Basis of Human Communication,” read at the University of Leeds, 1961, mimeo, available from C.L.R.U.
4. *The Oxford English Dictionary,* Oxford University Press, 1961.
5. Ullmann, S., *Semantics: an Introduction to the Science of Meaning,* Blackwell, Oxford, 1962.
6. Bally, C., “L'Arbitraire du Signe, Valeur et Signification,” *Le Français Moderne,* Vol. 8, 1940.
7. Lyons, J., *A Structural Theory of Semantics and its Application to some Lexical Sub-systems in the Vocabulary of Plato,* Ph.D. Thesis, University of Cambridge, 1961, published as *Structural Semantics,* Publications of the Philological Society, 20, Blackwell, Oxford, 1963.
8. Smith, C. J., A *Complete Collection of Synonyms and Antonyms,* London, 1867.
9. Firth, J. R. “Modes of Meaning,” in *Papers in Linguistics,* 1934-51, Oxford University Press, 1957.
10. *Webster's Dictionary of Synonyms,* G. C. Merriam & Co., Springfield, Mass., 1942.
11. Carnap, R., *Meaning and Necessity,* 2nd Ed., University of Chicago Press, 1956.
12. Quine, W. V., *Word and Object,* M.I.T. Press, Cambridge, Mass., 1960.
13. Naess, A., *Interpretation and Preciseness,* (Skrifter utgitt av Det Norske Videnskaps-Akademi i Oslo, Hist.-Filos. Klasse, 1953, No. 1), Oslo, 1953.
14. Austin, J. L., “A Plea for Excuses,” in *Philosophical Papers* (Ed. Urmson and Warnock), Oxford University Press, 1961.
15. Richards, I. A. and Gibson, C. M., *English through Pictures,* Pocket Books, New York, 1958.
16. Needham, R. M. and Parker-Rhodes, A. F., “The Theory of Clumps II,” 1960, mimeo, available from C.L.R.U.
    Needham, R. M., “The Theory of Clumps,” 1961, mimeo, available from C.L.R.U.
    Needham, R. M., *Research on Information Retrieval, Classification and Grouping, 1957-61,* Ph.D. Thesis, University of Cambridge, 1961.
    Needham, R. M., “A Method for using Computers in Information Classification,” *Information Processing 62: Proceedings of the IFIP Congress 1962,* North Holland, Amsterdam, 1963.
    Needham, R. M., "Applications of the Theory of Clumps," *Mechanical Translation* (this issue).
17. *The International Code of Signals, 1931,* British Edition, London, 1932.
18. Sparck Jones, K., *Synonymy and Semantic Classification,* Ph.D. Thesis, University of Cambridge, 1964.