

WHAT IS A THESAURUS?

by

Margaret Masterman

June, 1959

IntroductionPlan of paper and incompetence of present author.

Faced with the necessity of saying, in a finite space and in an extremely finite time, what I believe the thesaurus theory of language to be, I have decided on the following procedure:

Firstly, I give, in logical and mathematical terms, what I believe to be the abstract outlines of the theory. This account may sound abstract; but it is being currently put to practical use. That is to say, with its help, an actual thesaurus to be used for medium-scale Mechanical Translation tests, and consisting of specifications in terms of archeheads, heads, and syntax-markers, made upon words, is being constructed straight on to punched-cards. The cards are multiply-punched; a nuisance, but they have to be, since the thesaurus in question has 800 heads. There is also an engineering bottleneck about interpreting them; at present, if we wish to reproduce the pack, every reproduced card has to be written on by hand which makes the reproduction an arduous business; a business also which will become more and more arduous as the pack grows larger. If this interpreting difficulty can be overcome, however, we hope to be able to offer to reproduce this punched-card thesaurus mechanically, as we finish it, for any other M.T. group which is interested, so that, at last, repeatable, thesauric translations (or mistranslations) can be obtained.

I think the construction of an M.T. Thesaurus, Mark I, direct from the theory, instead of by effecting piecemeal changes in Roget's Thesaurus, probably constitutes a considerable step forward in our research.

In the second section of the paper, I do what can to elucidate the difficult notions of context, word, head, archehead, row, list as these are used in the theory. I do not think this section is either complete or satisfactory; partly because it rests heavily upon some C.L.R.U. Workpapers which I have written, which are also neither complete nor satisfactory. In order to avoid being mysterious, as well as incompetent, however, I have put it in as it stands. Any logician (e.g. Bar-Hillel) who will consent to read the material contributing to it, is extremely welcome to see this work in its present state; nothing but good can come to it from criticism and suggestion.

In the third section of the paper, I try to distinguish a natural thesaurus (such as Roget's) from a term-thesaurus (such as the C.L.R.U.'s Library Retrieval scheme), and each of these from a thesauric interlingua, (such as R. H. Richens' Nude). Each of these is characterised as being an incomplete version of the finite mathematical model of a thesaurus, given in Section I, - except that the Richens' interlingua has also a sentential sign system which enables Nude sentences to be reordered and reconstructed as grammatical sentences in an output language. This interlingual sign-system, when encoded in the programme, can be reinterpreted as a combinatory logic. It is evident, moreover, that some such sign-system must be superimposed on any thesaurus and the information which it gives carried unchanged through all the thesaurus-transformations of the translation programme, if a thesaurus programme is to produce translation into an output language. Thus, Bar-Hillel's allegation that I took up Combinatory Logic, as a linguistic analytic tool and then abandoned it again is incorrect; the bowler 'at's still there, guvner, if you 'ave a good look.

This section is also meant to deal with Bar-Hillel's criticism that "thesaurus" is currently being used in different senses. This criticism is dealt with by being acknowledged as correct.

The next Section asks in what ways, and to what extent, a language-thesaurus can be regarded as interlingual. We feel that we know a good deal more about this question than we did six months ago, through having now constructed a full-scale thesauric interlingua (Richens' Nude). This consists, currently of Nuda Italiana and Anglo-Nude. Nuda-Italiana covers 7,000 Italian chunks (estimated translating power, 35,000 words), and can be quasi-mechanically expanded ad lib by adding lists and completing rows. We are, however, not yet satisfied with it. It is being currently key-punched in a special code called Marcode, for further tests. The 48 elements of Nue-France also exist, but we are not yet developing it, since our urgent need is to construct a Nude of a non-romance language (e.g. Chinese): this will, we think, cause a new fashion to set in in Nudes, but will not, we hope, undermine the whole Nude schema.

In the final section of the paper, I open up the problem of the extent to which a sentence, in a text, can be considered as a sub-thesaurus. This section, like Section II, is incomplete, and unsatisfactory; I hope to take it up much more fully at a later date. It is so important, however, initially, to distinguish (as well as, I hope, finally to interrelate) the context lattice-structure of a sentence, which is a sub-thesaurus, from the sentential structure, which is not, that I have inserted this section, incomplete as it is, to try and make this one point clear.

We hope to issue a fuller report than this present one on the punched-card tests which we are doing and have done. We hope also to issue, though at a later date, a separate report on interlingual translation done with Nude. I should like to conclude this introduction by saying that we hope lastly and finally to issue a complete and authoritative volume, a sort of Principia Linguistica, or Basis Fundamentaque Linguae Metaphysicae, - devoted entirely to an exposition of the theory which will render obsolete all other expositions of the theory. I see no hope at all, however, of this being forthcoming, until an M.T. thesaurus (Mark N) survives large-scale testing on a really suitable machine.

Margaret Masterman
Cambridge Language Research Unit

1st June 1959

I. LOGICAL AND MATHEMATICAL ACCOUNT OP A THESAURUS
I. (a) GENERAL LOGICAL SPECIFICATION OF A THESAURUS
1. Basic Definition of a Thesaurus

A thesaurus is a language-system classified as a set of contexts.

(A context is further described below; it is a single use of a word.)

As new uses of words are continually being created in the language, the total set of contexts consist- ing of the thesaurus is therefore infinite.

2. Heads, lists and rows.

In order to introduce finiteness into the system, we therefore classify it non-exclusively in the following manner:

- i) The infinite set of contexts is mapped on to a finite set of heads. (Heads are further described below; they are the units of calculation of the thesaurus.) It is a prerequisite of the system that whereas the number of contexts continually increases in the language, the number of heads does not.
- ii) The contexts in each of these heads will fall into either a) lists, b) rows. (A list and a row are further described below. A list is a set of mutually exclusive contexts, such as "spade, hoe, rake"; which if used in combination have to be joined by "and"; a row is a set of quasi-synonymous contexts, such as "coward, faint-heart, poltroon", which can be used one after the other; if desired, in an indefinite string.)

3. Paragraphs and aspects.

- a) The heads are subdivided into paragraphs by means of syntax-markers. (A syntax marker is further described below; it is a very general concept, like the action of doing something, or the concept of causing somebody

to do something. Ideally, a syntax-marker specifies a paragraph in every head in a thesaurus. In fact, not every paragraph so specified will contain any contexts.

A paragraph can consist either of a set of rows in a head, or a set of lists; or of a set consisting of a combination of rows and lists. Such a set can have no members, (in which case, it is a vacuous set), one member or more than one member.

- b) The heads are cross-divided into aspects, by means of archeheads. (An archehead is further described below; it is a very general idea, such as that of "truth", "pleasure", "physical world".) A thesaurus-aspect consists, ideally, of a dimediate* division of the thesaurus (e.g. into "pleasing" and "non-pleasing" contexts). In actual fact, an archehead usually slices off an unequal but still substantial part of a thesaurus.

4. The resolving-power of a thesaurus.

It cannot be too much stressed that once the division into heads, paragraphs, rows, lists and aspects has been effected, the contexts of the thesaurus are not further subdivided. This limit of the power of the thesaurus to distinguish contexts is called the limit of the resolving-power of the thesaurus; and it is the great limitation on the practical value of the theory. Thus, the thesaurus theory of language does not, as some think, solve all possible linguistic problems; it does, however, successfully distinguish a great many contexts in language in spite of the fact that none of these contexts can be defined.

To find the practical limits of the resolving-power of any thesaurus should thus be the first object of any thesaurus research.

* A dimediate division is a binary chop.

I (b) A FINITE MATHEMATICAL MODEL OF A THESAURUS1. Procedure of conflating two oriented partially-ordered sets.

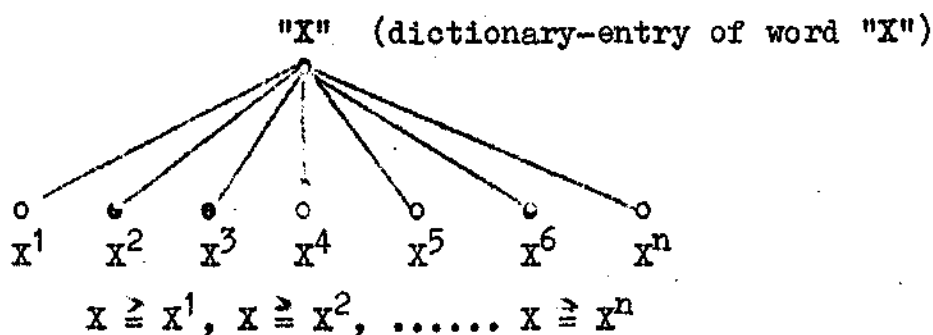
When a finite mathematical model is made of a thesaurus, the non-exclusive classification generates a partially-ordered set. By adding a single point of origin at the top of the classification, this set can be made into an oriented partially-ordered set, though it is not a tree.

It must be remembered, however, that, if it is to have an empirical foundation, a thesaurus of contexts must also be a language of words. An actual thesaurus, therefore, is a double system. It consists: i) of context-specifications made in terms of archeheads, heads, syntax-markers and list-numbers; and it also consists ii) of sets of context-specifications which are uses of words. Now, a case will be made, in the next section, for defining also as an inclusion relation the relation between a dictionary-entry for a word, (that is, its mention, in heavy type, or in inverted commas, in the list of words which are mentioned in the dictionary) and each of the individual contexts of that word (that is, each of the definitions given, with or without examples, of its uses, and which occur under the word-entry in the dictionary). In the next section, it will be argued in detail that such a relation would generate a partially ordered set but for the fact that, owing to the same sign, or a different sign, being used indiscriminately both for the dictionary-mention of the word and for one or any or any number of its uses, the axioms of a partially-ordered set can never be proved of it. This is my way of approaching the fundamental problem of the "wobble of semantic concepts" which Bar-Hillel correctly brought up in Area 6 of I.C.S.I., and which unless some special device is developed to deal with it, prevents any logical relations between semantic units ever being provable. Now, a thesaurus is precisely a device for steadying this wobble of semantic signs; that is one way of saying what it is; and the device which it uses is to define, not the semantic signs themselves, nor their uses, but the thesaurus positions in which these uses occur. The same word-sign, therefore, i.e. the same conceptual sign, i.e. the same semantic sign, occurs in the thesaurus as many times as it has distinguishable contexts; a word like "in" which has, say, 200 contexts in English,

will therefore occur in the thesaurus 200 times. Thus, the theoretical objection to arguing on the basis that the relation between a dictionary-mention of a word and its set of contexts is an inclusion-relation disappears as soon as these contexts are mapped on to a thesaurus.

In this section we assume, therefore, what in the next section we argue that we can never prove; namely, that the relation between a dictionary mention of a word and the items of its entry itself generates an oriented partially-ordered set.

Fig.I Oriented partially-ordered set consisting of the dictionary entry of a word.



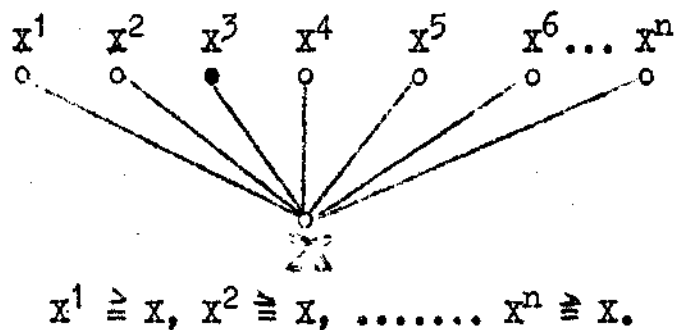
But now, we have to notice an important logical fact. This is, that a use of a word as it occurs in an actual text (what is, when it is actually used, not mentioned) is logically different from the heavy-leaded type mention of the word when it is inserted as an item of a dictionary. For the word as it occurs "in context", as we say, - i.e. in an actual text in the language, - by no means includes all the set of its own contexts. On the contrary, the sign of the word there stands for one and only one of its contexts; it therefore stands also for a context-specification of this use made in terms of archeheads, heads, syntax-markers and list-numbers (see above).

This assertion requires a single proviso: which is that in a text (as opposed to in a language) the set of archeheads, heads, syntax-markers and list-numbers needed to make the context specifications of the constituent words will be a subset of the set consisting of the total thesaurus; namely, that subset which is needed to specify the contexts of the actual text. Thus, the contexts used in

any text (or any sentence) in a language will be a sub-language-system, consisting of a sub-thesaurus (see Section V).

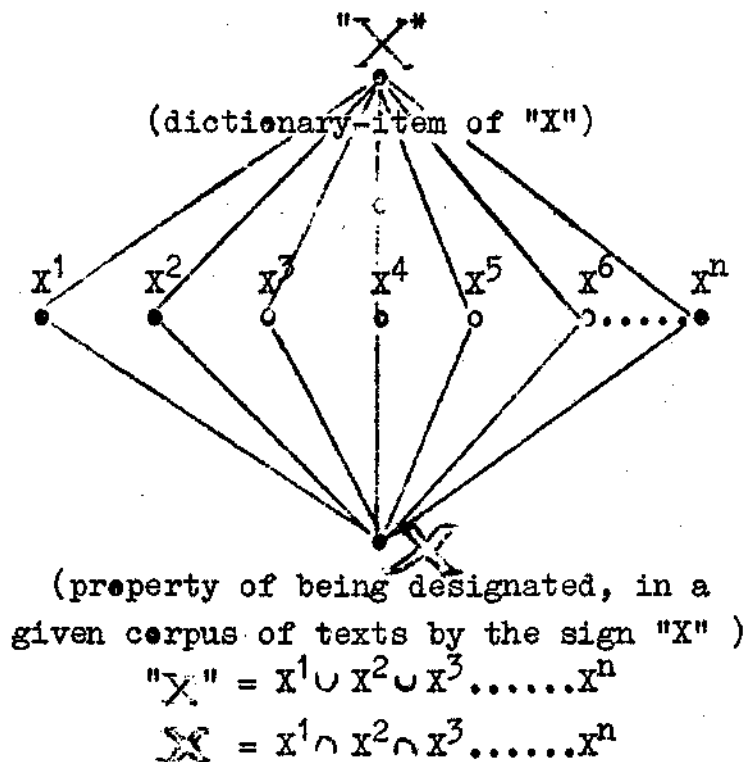
This fact alters the nature of the mathematical model which it was proposed to make of a thesaurus. For the word, as it is used in all the texts of the language (as opposed to the word as it is mentioned in the dictionary), now consists of that which is in common between all the context-specifications which occur in all the texts; these context-specifications being in terms of archeheads, heads, syntax-markers and list-numbers (see above). Because all that is in common between all these text-specifications, so made, is the empirical fact that all of them can be satisfactorily denoted in the language, by the sign for that one word. When it is inserted into a thesaurus, therefore, as opposed to when it is inserted as part of a dictionary, the oriented partially-ordered set consisting of the set of uses of a word becomes inverted, (i.e. it has to be replaced by its dual), because the inclusion relation becomes reversed.

Fig. II. Oriented partially-ordered set, dual of the set given above, consisting of the dictionary-entry of a word, consisting of the relation between the word-sign and the total set of its possible contexts, as appearing in texts.



It follows, if partially-ordered set II is the dual of partially-ordered set I, that they can be combined into one partially-ordered set. It is easy to see intuitively that the partially-ordered set so formed is the "spindle-lattice" of $n + 2$ elements.

Fig. III - Spindle-lattice formed by conflating the two partially-ordered sets given above:

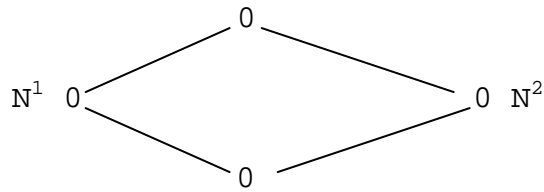


It may be a help to see that the interpretation of the meet and join relations which is here made, has an analogy with the interpretation of a Boolean lattice which is given when the meet and join relations are imagined to hold between numbers. Thus, in a 4-element Boolean lattice of which the side-elements are numbers, N^1 and N^2 , the join of these two numbers will be their least Common Multiple, and the meet of the same two numbers will be their Highest Common Factor. Analogously, in the interpretation which we are making, the join of the two contexts of a word, C^1 and C^2 , will be the dictionary-entry listing both of them, and the meet will be any property which is in common between them; in this case, the property of being denotable by the sign of the same word.

This analogy is illustrated diagrammatically below:

I. Numerical Case

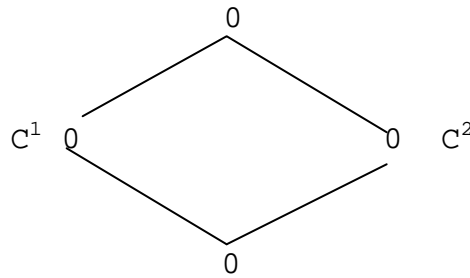
L.C.M. of N^1 and $N^2 = N^1 \cup N^2$



H.C.F. of N^1 and $N^2 = N^1 \cap N^2$

II. Word Case

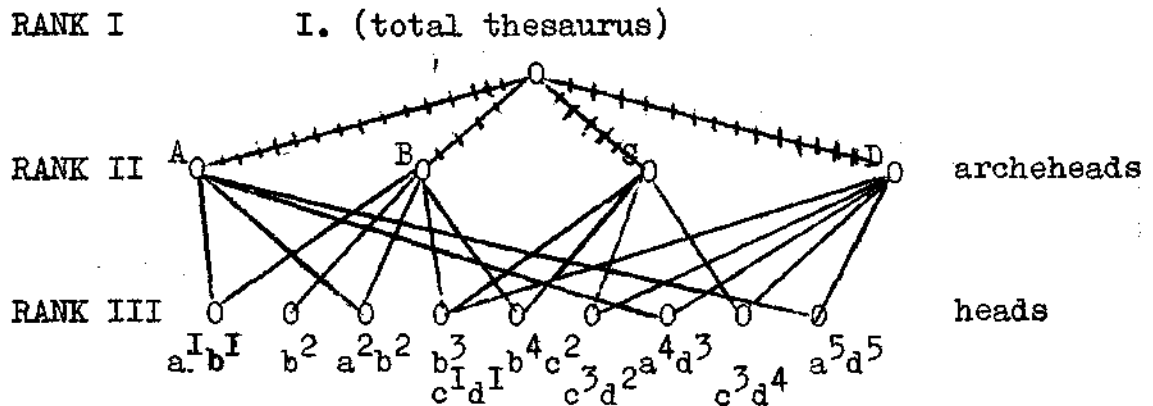
Dictionary-entry of C^1 and $C^2 = C^1 \cup C^2$



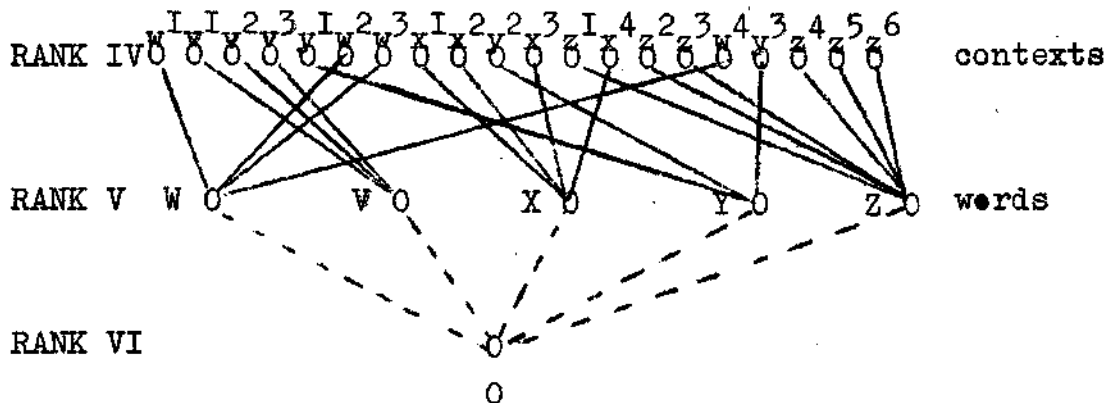
Property of there being the same word sign
for both C^1 and $C^2 = C^1 \cap C^2$.

To return now to the thesaurus model. If it be granted that partially-ordered set I and partially-ordered set II can be conflated, without empirical or mathematical harm, to form the second lattice, it will be no empirical or mathematical surprise to find that, on the larger scale also, two oriented partially-ordered sets can be conflated with one another to form a figure which has a tendency to become a lattice.

For, whereas the total archeheads and heads of the thesaurus form an oriented partially-ordered set of this form:



The words and their contexts in the thesaurus, (not in the dictionary) form an oriented partially-ordered set of this form:



(property of being a word in a language)

2. Procedure for converting the conflation given under 1 into a finite lattice.

Mathematically, it will be easily seen that there is no great difficulty in converting the figure, given above, into a finite lattice. If it is not a lattice already, all that is required, is to make it one, by adding, vacuously, extra context-points wherever sufficient meets and joins do not occur. If, upon test, an extra rank begins to show up below the word-sign rank, and corresponding to the archeheads, it will probably be possible, with a minimum of adjustment, to embed this thesaurus in the lattice A^3_5 , (attached to the end of this section) which is the cube (A^3) of the spindle of 5 elements (A^5). Of course, if any of the vacuous context-points turn out to "make sense" in the language, then word-uses or phrase-uses can be appointed to them in the thesaurus, and, in consequence, they will no longer be vacuous.

Empirically, however, - however desirable it may be mathematically, - there seems to be grave objection to this procedure. For even if we ignore the difficulty, (which is discussed in Section IV) of determining what we have been meaning throughout by "language", it yet seems at first sight as though there is another objection in that we have been conflating systems made with two inclusion-relations; namely, i) the theoretic classifying-relation between heads, archeheads and contexts, and ii) the linguistic relation between a word and its contexts. If we look at this matter logically, however, (that is, neither merely mathematically nor merely empirically) it seems to me that the situation is all right. For even if we get at the points, in the first place, by employing two different procedures, (i.e. by classifying the contexts, in the librarian's manner, by means of archeheads and heads, whereas we deploy the contexts of a word, in the dictionary-maker's manner, by writing the sign for it under every appropriate head), yet logically speaking, we have only one inclusion-relation which holds throughout all the ranks of our thesaurus. For the heads, as

well as having special names of their own, can also be specified, as indeed they are in the lattice-like figure, as being intersections of archeheads. Similarly, the contexts on the rank lower down could be specified not merely in terms of the units of the rank immediately higher up, i.e. of the heads, but also as intersections of heads and archeheads. And as we have already seen, at the rank lower down still, word signs can be seen as intersections of their contexts, and therefore, specifiable also in terms of intersections of archeheads and heads.

It may be asked whether there is any difference, on this procedure, between a good and a bad thesaurus-lattice. To this, it may be replied that the second object of any thesaurus research, should be to discover how many vacuous context-points remain vacuous (i.e. cannot have any word-uses or phrase-uses attached to them) when any given thesaurus is converted into a lattice. On the ordinary canons of scientific simplicity, the more vacuous context-points have to be created, the less the thesaurus, in its natural state, is like a lattice. Conversely, if (as has been found), very few such points have to be created, then we can say in the ordinary scientific manner, Language has a tendency to be a lattice,*

* Eighteen months ago, the Cambridge Language Research Unit was visited by the director of a well-know British Computer laboratory, who was himself very interested in the philosophic "processing" of language. On the 'phone, before he arrived, he announced that his point of view was, "If language isn't a lattice, it had better be." Sometime later, after examining the C.L.R.U. evidence for the lattice-like-ness of a language, and what could be done with a lattice-model of a thesaurus, he said mournfully, and in a quite different tone, "Yes, it's a lattice; but it's bloody large."

3. Syntax-markers: the procedure of forming the direct product of the syntax-lattice and the thesaurus lattice.

The argument up to this point, if it be granted, has established that a finite lattice-model can be made of a thesaurus. It has only established this fact, however, rather trivially, since the classificatory principle of A^3_5 is still crude. It is crude empirically since it embodies, at the start, only the amount of classification which the thesaurus compiler can initially make when constructing a thesaurus. Thus the initial classification of what one finds in "language", is into archeheads, heads, syntax-markers, list-numbers and words. It is also crude mathematically, since the lattice A^3_5 , splendid as it looks when drawn out diagrammatically, is founded only upon the spindle of five elements; and, in this field, a spindle is of all lattices the one not to have if possible, since it represents merely an unordered set of concepts with a common join and meet.

Two things are needed to give more "depth" to the model; firstly, the structure of the syntax-markers, which have been left out of the model entirely so far; secondly, an unambiguous procedure for transforming A^3_5 which, on the one hand, will be empirically meaningful, and on the other hand, will give a lattice of a richer kind.

Let us consider the syntax-markers first. Two cases only are empirically possible for these: i) that they are similar in function to the archeheads, being, in fact, merely extra archeheads which it has been convenient, to somebody, for some reason, to call "syntax-markers"; ii) they are different in function from archeheads, as asserted in I.3; in which case this difference in function must be reflected in the model. Now, the only empirical difference allowable, in terms of the model, will have

* Of these, using Roget's Thesaurus as an example of "language", the archeheads, (in so far as they exist) are to be found in the Chapter of Contents, though they usually represent somewhat artificial concepts; some of the heads themselves, though not all, are arbitrary; the syntax-markers, noun, verb, adjective and adverb, are not interlingual; finally, instances of every length of language segment, from morpheme to sentence, are to be found among the words. (See later Sections II, III & IV.)

to be that whereas each archehead acts independently of all the others, picking out its own substantial subset of the total set of the thesaurus, the syntax-markers act in combination, to give a common paragraph-pattern to every head. And this means that the total set of syntax-markers will form their own syntax-lattice; this lattice, taken by itself and in isolation, giving the pattern which will recur in every head.

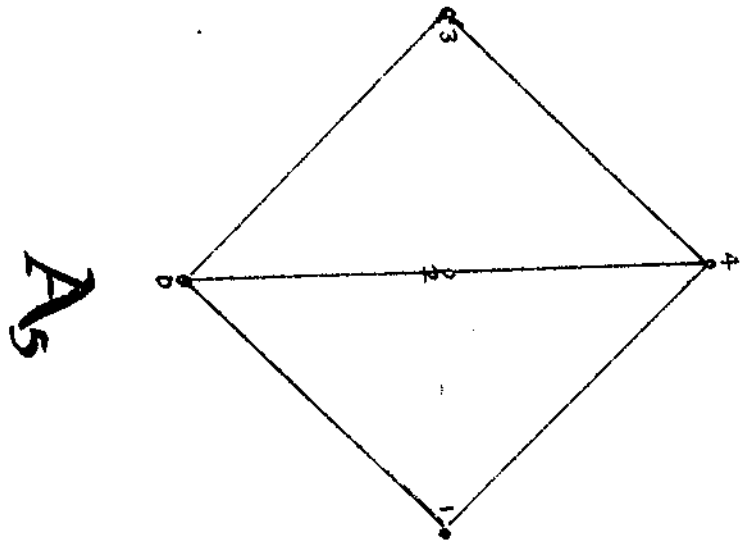
It is thus vital, for the well-being of the theory, that the lattice consisting of the total set of syntax-markers should not itself, (as indeed it tends to do) form a spindle. For this fact implies that the set of syntax-markers, like the set of archeheads, is unordered; in which case, the markers are merely archeheads. If, however, without damage to the empirical facts, the syntax-markers can be classified into mutually exclusive subsets*, then the situation is improved to that extent; for the syntax-lattice will then be a spindle of spindles. And any further ordering principle which can be discovered among the syntax-markers will improve the mathematical situation still further; since it will further "de-spindle" the paragraph-pattern of the heads. But such an ordering principle must be discovered, not invented; for the allowable head-pattern for any language, is empirically "tight" in that much more than the set of heads, it is an agreed and known thing. Moreover, if it is to pay its rent in the model, it must be constant throughout all the heads, though sometimes with vacuous elements. For if no regularity of paragraph-pattern is observable in the heads, then it is clear that, as when the syntax-lattice was a spindle, the syntax-markers are again only acting as archeheads. The former betrays itself in the model: There will be a huge initial paragraph pattern, large parts of which will be missing in each head.

Thus the construction of the syntax-lattice is fraught with hazards, though the experimental reward for constructing it correctly is very great⁺. The procedure for incor-

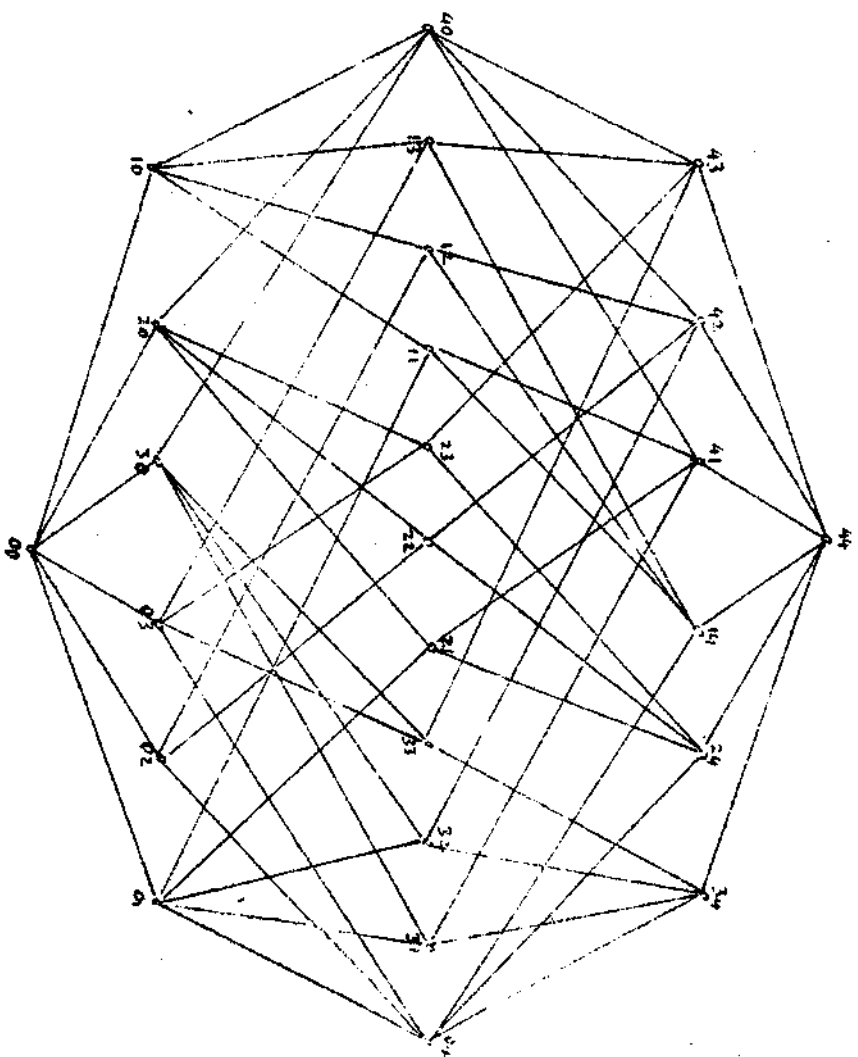
* They are so classified in the Interlingua Lattice; but not in the logically more primitive Interlingua Nude (see Sect. II)

+ See Section V.

The Initial Powers of the Spindle of Five



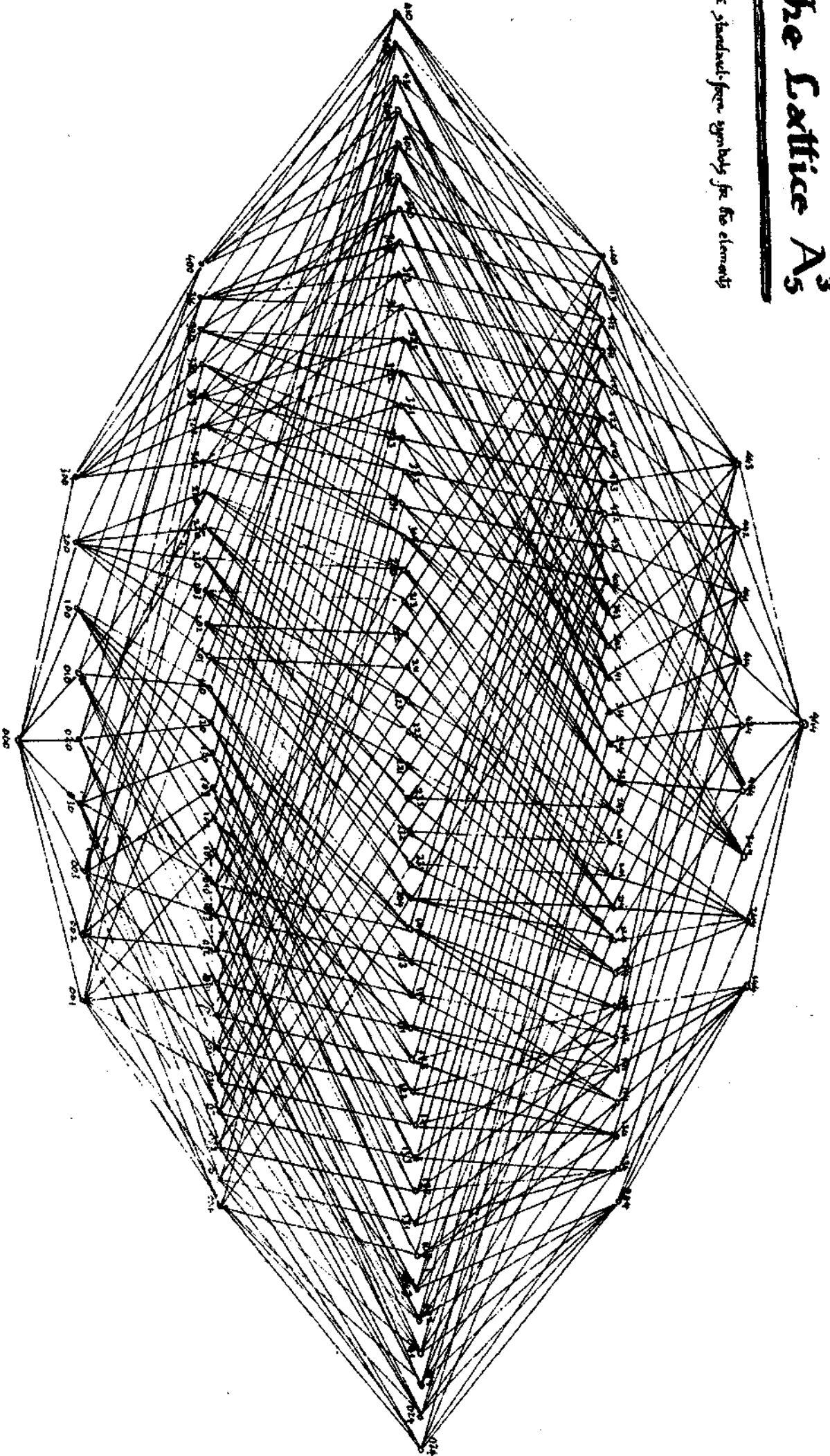
A_5



A_5^2

The Lattice A_3

with standard form symbols for the elements

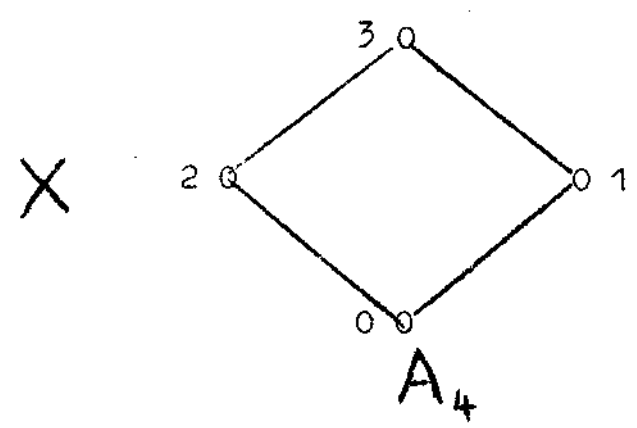
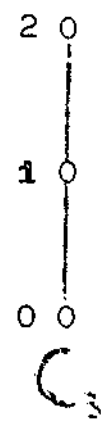


porating it in the model, however, is unambiguous: a direct product is formed of thesaurus-lattice and syntax-lattice, this product forming the total lattice of the language. This total lattice can be computed but not displayed, since it is quite out of the question to present in diagram form the direct product of a spindle of spindles with A^3_5 . The principle of forming such a direct product, however, can be easily shown; it is always exemplified by the very elegant operation of multiplying the Boolean lattice of 4 elements by the chain of 3.

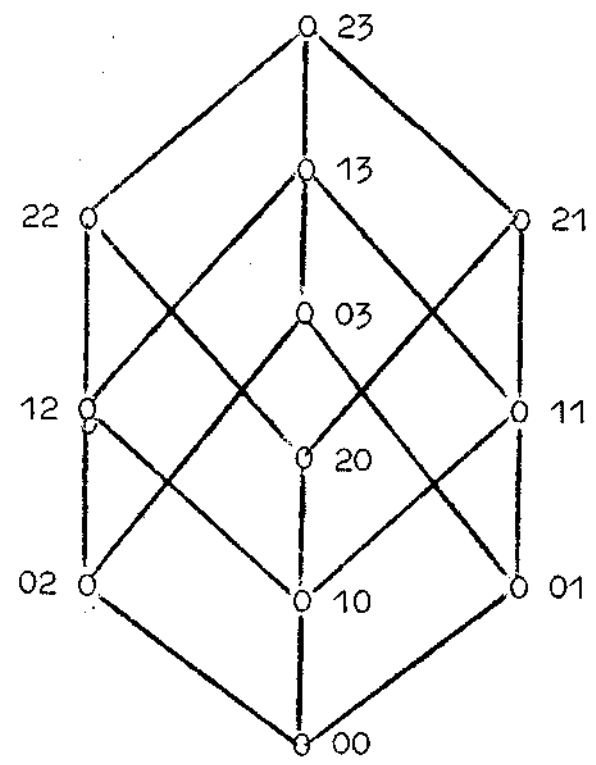
(See Diagram Overleaf)

And a sample syntax-lattice, like a simple direct product, can be constructed. But in even suggesting that it should be constructed, I am putting the logical cart before the logical horse. For it is precisely the set of lattice-operations which I am about to specify which are designed to enable thesaurus-makers objectively to re-structure (which means also, by the nature of the case, to "de-spindle") both the syntax-lattice and the thesaurus. Until we have the data which these operations are designed to give, it is not much use imagining a thesaurus-lattice except as embedded in A^3_5 , or a syntax-lattice except as a spindle of sub-spindles, the points on each sub-spindle carrying a mutually exclusive subset of syntax-markers. The total sets of syntax-markers which we have been able to construct are not nearly sufficient, by themselves, to give grammatical or syntactical systems for any language. They are, however, interlingually indispensable as output assisting signals, which can be picked up by the monolingual programme for constructing the grammar of the output text, or even the semantic part of the output-finding procedure. As assistances to grammar, they are very useful indeed; for since they are semanticised, rather than formalised, they can straightforwardly operate on, and be operated on by, the other semantic units of the thesaurus. Thus they render amenable to processing the typical situation which arises when it comes to the interlingual treatment of grammar and syntax; the situation, that is, where information which is grammatically conveyed in one language, is conveyed by non-grammatical, i.e. by semantic means, in the next.*

* See the companion paper in this volume by Martin Kay.



=



$A_4 C_3$

Lattice Operations on a Thesaurus.

i. The Translation or Retrieval algorithm.

This is the process of discovering from a specification, given as a set of heads, an element of a given set with as nearly as possible the specified heads. This is exemplified by the procedure used in the rendering of "Agricola incurvo....", which is attached to the paper in this volume by Bastin and Needham. There, however, it is only applied to the semantic thesaurus, not to the language lattice as a whole⁽¹⁾. It is also used in the Cambridge Language Research Unit's Library Retrieval System⁽²⁾ where it is refined for practical purposes by a theorem⁽³⁾ which enables an easier procedure to be used to the same effect. Further versions of this algorithm will undoubtedly be needed (see under "metrics" below).

ii. Compacting and expanding the Thesaurus.

This is the process of making some of the heads more inclusive or more detailed, in order to affect the distinctions made by the heads or to change the number of heads used. An example of this process is described by M. Shaw⁽⁴⁾ when it was found necessary for coding purposes to have only 800 heads rather than 1,000.

iii. Embedding the total lattice in other lattices.

This again is an operation performed, primarily for coding purposes; it depends essentially on the theorem that any lattice can be embedded in a Boolean lattice. From this it is possible to derive a number of theorems and methods for handling thesauric data economically⁽⁵⁾. However, the process also throws some light on the logical structure of the whole thesaurus.

1. Masterman, Needham, Spärck Jones. As a translation algorithm, it is given in Masterman, Potentialities of a Mechanical Thesaurus (M.I.T., 1956) and in The Analogy between Mechanical Translation and Library Retrieval.

2. Joyce & Needham, Amer. Doc, 1958.

3. Needham, "Research Note on a Property of Finite Lattices," C.L.R.U., 1958.

4. "Compacting Roget's Thesaurus", C.L.R.U., 1958.

5. Parker-Rhodes & Needham, "Computation Methods in Lattice Theory", submitted to Camb. Phil. Soc.; also a "Reduction Method for Non-arithmetic Data", I.C.I.P., Paris, 1959.

iv Extracting and performing lattice operations on
sentential sublattices. (See Section V:)

v. Criteria for nearness of fit.

It is possible to regard a lattice as a metric space in several ways, and as having a non-triangular pseudo-metric in many others. To do this, in practice, is extremely difficult, though the task is not, we still think, an impossible one. The obvious criterion of thesaurus-lattice distance is "number of heads in common"; for instance, if there are 10 words in common between the head Truth and the head Evidence, 7 words in common between the head Existence and the head Truth, and 3 words in common between the head Existence and the head Evidence, it might be thought that, by counting the words in common, we could establish a measure of their relative nearness. Consider, however, the possible complication: Existence might have 50 words in it, Evidence 70, Truth 110; this, already complicates the issue considerably. Then there are the further questions of aspect and paragraph-distinction; are similarities in those respects to contribute to "nearness"? One such is embodied in the Translation algorithm above, and research is in progress on the selection of the most appropriate one for translation purposes. For example, it is necessary to be able to say whether a word with heads, A,B,C,D,C, is nearer to a specification B,C,D,F, than a word with heads C,D,F,G. The remaining two kinds of operation are concerned with testing a thesaurus rather than using it.

vi Finding the resolving power.

This consists of discovering what sets of words have exactly (or once a metric has been agreed, nearly) the same head descriptions. The closeness of the intuitive relation between these words is a test of the effectiveness of the thesaurus.

5, The impossibility of fully axiomatising any finite lattice-model of a thesaurus.

A thesaurus is an abstract language-system; and it deals with logically primitive language. That this is so can be seen at once as soon as one envisages the head-signs as logically homogenous ideographs. The words (to distinguish them from the heads), could then be written in an alphabetic script. But what kind of sign are we then to have for the syntax-markers? What kind of sign, also, for the archeheads? Different coloured ideographs, perhaps; or, ideographs enclosed in squares for the syntax-markers, and ideographs enclosed in triangles for the archeheads.

A thesaurus is an abstract language-system, and it deals with logically primitive language. It therefore looks, at first sight, as though it were formalisable; as though the next thing to do is to get an axiomatic presentation of it.

That it is logically impossible to get such a formalisation however, becomes apparent as soon as one begins to think what it would really be like. Imagine a thesaurus, for instance, typographically set out so that i) all the head-signs were pictorial ideographs, ii) the archeheads were similarly ideographs, each however enclosed in a triangle; and iii) the syntax-operators were similarly ideographs, again, each being enclosed, however, in a square. Would it not be vital to the operation of the thesaurus to be able both to distinguish and to recognise the ideographs? To know, for instance, that the ideographic sign for "Truth", (say, a moon exactly mirrored in a pond) occurred also in the archehead "Actuality", which will be a moon mirrored in a pond, and enclosed in a triangle?

Moreover, imagine such a system "mathematicised"; i.e. that is re-represented in a different script; that is, with its ideographs replaced by various alphabets (you would need several), and the triangular and square enclosures respectively by braces and square brackets? What have you done, when you have effected this substitution,

except replace ideographs by other ideographs? Are not A, B, C, D ideographs? Are not brackets ideographs? And is E not as important in the alphabetic as in the pictorial case, to know that A is not B, and B is not C; to distinguish (A), or [A] not only from A, but also from B, or (B), or [B]? There could be no better case than this for bringing home the truth, - which all logicians in their heart of hearts really know - that there are required a host of conventions about the meaningfulness and distinguishableness of ideographic symbols before any ideographic system can be formalised at all. In a C.L.R.U. Workpaper issued in 1957*, I wrote:..."What we are analysing, in analysing the set of uses of a word, is the situation at the foundations of all symbolism, where the normal logical sign-substitution conventions cannot be presumed to hold. Because exactly what we are studying is, 'How do they come to hold?'... By mathematical convention, then, if not by mathematical assertion, variables have names..." (In fact, a mathematical language which consisted of nothing but variables, like a thesaurus, would be logically equivalent to St. Augustine's language, which consisted of nothing but names). A mathematical variable has meaningfulness and distinguishableness in a system because it has the following three characteristics:

- i) It is a name for the whole range of its values; we learn a lot about these values by naming the name. The traditional algebraic variables x and y , stand for numerals; the traditional variables p , q , r , stand for statements? and so on.
- ii) It has a type: it occurs in systems which have other signs which are not variables, (e.g. the arithmetical signs, or the propositional constants) from which it can be distinguished by its form.
- iii) It has context: that is to say, by operating with one or more substitution-rules, a further symbol giving a concept with a single meaning, can be substituted for the variable.

* In the paper, I took the combinator-rules of a combinatory logic; and by progressively removing naming-power and distinguishability from the symbols, produced a situation where no one could tell what was happening at all.

Now as soon as we operate with the heads of a thesaurus, we operate with variables from which the second characteristic has been removed*. The result of this is that the first and third characteristics, namely that a mathematical symbol is a name, and that it has context, acquire an exceptional prominence in the system, and that whatever system of mathematical symbols you use. Why, then, give yourself a great effort of memory learning new names, when names already approximately existing in your language, and the meaningfulness and distinguishableness of which you know a good deal about already, will perfectly well do?

Another, general way, of putting this argument is by saying that any procedure for replacing the head-signs by other signs will be logically circular. For in the model, as soon as we replace the archehead or head specifications by formal symbols, we can only distinguish them one from another by lattice-position. But we can only assign to them lattice-position if we can already distinguish them from one another. In making this model, Language, (philosophic English, L_1) is being used to construct a Language (the heads, archeheads, markers, list-numbers of the thesaurus and the rules for operating them, L_2) to analyse Language (the words and contexts of a natural language, L_3). Every attempt is made, when doing this analysis, to keep L_1 , L_2 and L_3 distinct from one another. But there comes a point, especially when attempting formalisation, beyond which the distinction between the three goes bad on you; and then the frontier-point in determining the foundations of symbolism has been reached. Beyond that point, variable and value, variable and constant, mathematical variable and linguistic variable, sign and meta-sign - it's all one: all you can do is come up again, to the same semantic barrier, by going another way.

In our thesaurus, in order to avoid the use of ideographs, archeheads are in large upper-case letters and followed by a shriek (e.g. TRUE!), heads are in small upper-case letters with a capital, (e.g. EVIDENCE, TRUTH); words are in ordinary lower-case letters (e.g. actual, true); and syntax-markers are hyphenated and in italics, (e.g. fact, concrete-object).

* In the model, the heads, etc, can of course be distinguished from the lattice-connectives. To that extent, but only to that extent, the system is formalisable.

II. CONTEXTS, WORDS, HEADS, ARCHEHEADS, ROWS, LISTS.

1. Contexts.

It is evident that if we wish to come to a decision as to the extent to which thesaurus-theory has an empirical foundation, the vital notion to examine is that of context.

Having said this, I propose now to examine it, not concretely but abstractly; because in the course of examining it abstractly, it will become clear how very many obstacles there are to examining it concretely. Roughly, if a language were merely a large set of texts, there would be no such difficulty; research with computers would show to what extent these could be objectively divided up by using linguistic methods, and into how small slices; a list of the slices of appropriate size, (i.e. morphemes, rather than phonemes,) would be the contexts. Actually, however, language is not like that. Firstly, nobody knows how large a number of texts, and what texts, would be required for these to constitute a true sample. Secondly, we have to know quite a lot about any language, both as to how it functions and to what it means in order to give the computer workable instructions as to how to slice up the text. So even if we wish to be 100% empirical - "to go by the facts and nothing but the facts" - we find that a leap of the creative intellect is at present in fact needed to arrive at a purely empirical notion of collocation, or context. And that being so, there is everything to be said, for using to the full, in an essentially general situation, the human capacity to think abstractly.*

The argument which follows comes from the same 1957 unpublished C.L.R.U. workpaper, Fans and Heads, from which I quoted in the previous section. The argument on the difficulty of defining a word, however, comes from a published paper (Masterman, Words, Proceedings of the Aristotelian Society. 1952-53).

* i.e. if we have to take a creative leap, in any case, let it not be a naive one; let us do our best to turn it into an informal theoretic step.

The philosopher Wittgenstein, in his book Philosophical Investigations (1953), compared the set of uses of a word to the set of ways in which one can see a gestalt figure. At the time, he was primarily investigating, not the concept consisting of the set of uses of a word, but the concept considered as a gestalt given by perception; of how it itself, (considered as an actual percept, given by experience, and also, by extension, as a picture, or an image) can be affected by environment, or context.

That this is so can be shown from the following passage:

II, xi, p. 193^e "I contemplate a face, and then suddenly notice its likeness to another face. I see that it has not changed; and yet I see it differently. I call this experience 'noticing an aspect'.
"Its causes are of interest to psychologists.
"We are interested in the concept and in its place among the concepts of experience..."

That Wittgenstein thinks, however, that there is an analogy (as well as a contrast) between the way in which context affects the "seeing" of a percept, and the way in which context affects the meaning of a word can be shown by the following passage:

II, xi, p. 210^e "I can imagine some arbitrary cipher -
this, for instance: } to be a strictly correct
letter of some } foreign alphabet. Or
again, to be a } faultily written one,
and faulty in this } way or that: for example, it
might be slap- } dash, or typical childish

awkwardness, or like the flourishes in a legal document. It could deviate from the correctly written letter in a number of ways. - And I can see it in various aspects, according to the fiction I surround it with. And here there is a close kinship with experiencing the meaning of a word."

So, a context, seen abstractly, is an intuitively given thing; it is a fiction. You can experience it; you can describe it up to a point; but you can't define it; seeing a context is like seeing a visual analogy.

Without pausing to see whether this very general idea of a context is right or not, let us now explicitly examine the set of uses of a word. For if my first point is that contexts must first be "seen" before they can be "found", and that there is no deeper analysis which we can at present make of this "seeing", my second point is that any attempt to define

mathematically and in vacuo the relations between concepts fails because of the looseness of fit between sign and signified.

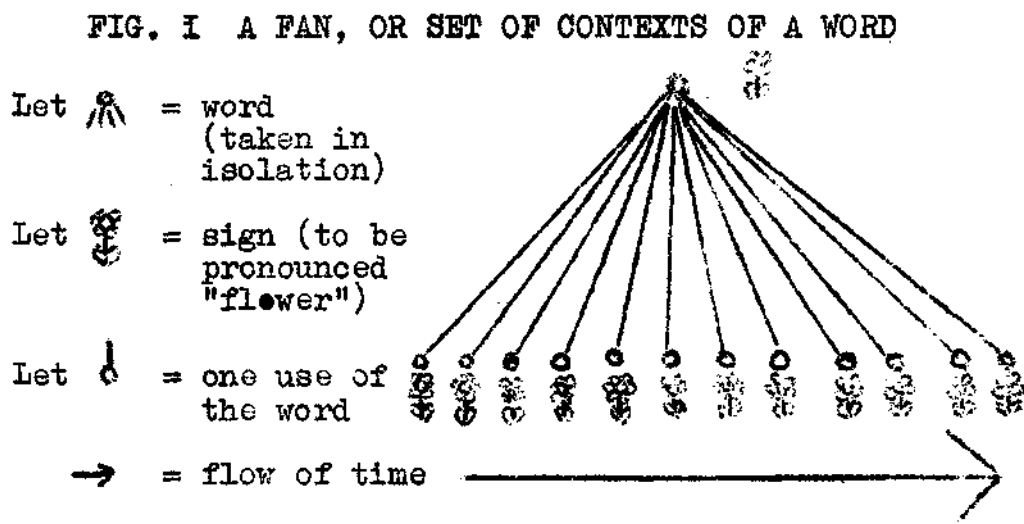
Let us take a word, its sign, and its set of uses.

Let us, in the simplest case, relate these to reality; about which we shall say no more than, since it develops in time, the uses of the word must also develop in time.

Let us denote the word by a point; its sign by an ideograph*; and its set of uses, .. all linking on to reality at unknown but different points, but all radiating out from the original point denoting the word, because they are all the set of uses of that same word, by a set of spokes radiating from that point.

Let us call the logical unit so constructed, a FAN.

I will give the essential idea of such a fan:



From this figure several facts can be made clear. The first is that, however many uses the word may have (however many spokes the fan may have) they will always be marked with the same sign. But it does not follow from this that all the uses of the word mean the same thing; that they all have the same meaning in use. It follows from this, on the contrary, that there is no one-one correspondence between sign and signified of the kind which logicians have always considered as an essential prerequisite for the construction of a mathematical theory of language; and that therefore a

* The case for denoting language-signs by ideographs, as soon one is searching for logical primitiveness in language, is given by implication in Section I. Roughly, ideographs look logically homogeneous; (they aren't, but they look it) whereas language-signs given in alphabetic script have a false precision; we see them as though they were in different parts of speech.

fresh type of mathematical construction must be envisaged; one, that is, which allows for a looser type of "fit" between sign and signified. As shown in the key to Fig. I, the point of origin of the fan will be the dictionary-entry of the word - that is, the word taken in isolation, and in the totality of its uses, without it being initially clear that these uses are. Each of the spokes of the fan will then give one actual use of the word; it being presumed that each actual use of the word can be, theoretically at any rate, taken in isolation.

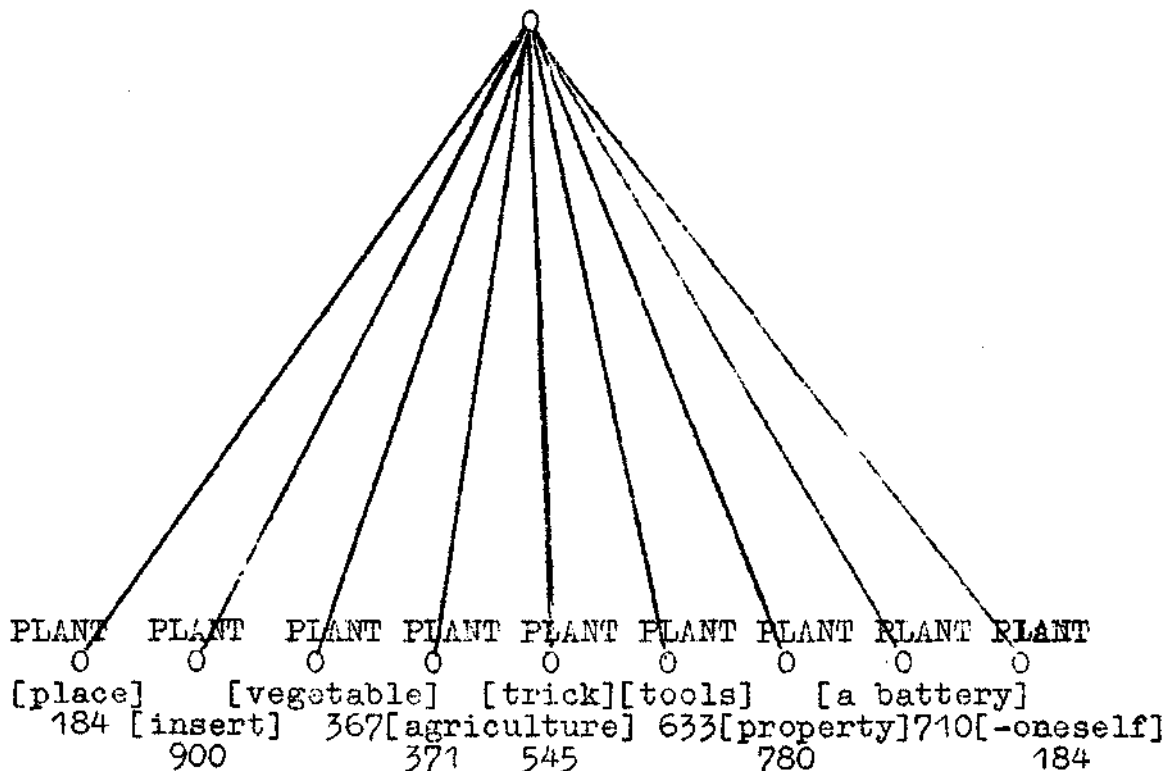
So we know that the fan has a point of origin, and spokes; we also know that it has a sign. What we do not know, finally or definitively, is how many spokes any fan has. This fact is indicated, in Fig. I, by the presence of the arrow indicating the flow of time. For at any moment, as we know, any new use of any word may be being created; and there will be no formal marker of this fact, since, as we have seen, all the uses of the word will be marked by the same sign. Now, we can lay down, from our general knowledge of the language, that there will be a finite number of spokes in any fan, provided that the total dictionary-entry only takes account of the evolution of the language from a time T_1 , (the approximate date when we first heard of the language) to a point T_2 (the present time). For if we assume that any new use of any word in any language takes a finite stretch of time to establish itself, and that every language had a definite beginning in time, then it will follow that no word in any language has an infinite number of uses; that the number of spokes in any fan will be countable and finite. On the other hand, if we consider the totality of use in the language - that is the use of a fan, for any T_2 then it will be clear that the number of uses of any word, though still finite until the language has lasted for an infinite stretch of time, is nevertheless tending to get larger and larger. This gives the number of spokes in any fan the property of a Brouwerian infinity that is, of getting progressively larger and larger; which means in its turn that any theorem about the language will have to be proved for the infinite case, as well as for the finite case.

That the figure which we have constructed of the set of uses of a word is not wholly fanciful, say, from the point of view of linguists, may be seen by constructing a similar figure, using the same conventions, from the entry "PLANT", taken from the cross-reference dictionary of Roget's Thesaurus.

FIG. II

PLANT

(alphabetically given dictionary entry)



I now want to establish, for any fan, two laws. I will call these The Fan Law and The Context Law.

Let us now consider the fan law of any fan. It will consist of an amplified definition of a fan.

FAN LAW OF ANY FAN A fan is a formal construction such that

1. it has a sign, which we will call the sign of the fan;
2. it has a point of origin, which we will call the hinge of the fan;
3. it has an unknown number of connections between its point of origin and a row of other points; we will call these connections the spokes of the fan, and the row of points so connected with the point of origin the row of points of the fan.

4. in the case of any spoke, we shall call the relation which connects the point of origin with the row of points the determination relation of the fan.





This fourth clause of the fan law at once brings up the question as to whether we can say more, in the fan law, about the determination-relation of the fan. And here we have to ask ourselves: "What would it be like to know more of this relation?" To this it may be answered that we know already (in the sense that we have assumed already) that it is a single relation; can we similarly, (and in the same sense) know any more about it?

It seems to me evident that we can, once we admit of information taken from the context law of any fan as being relevant to the formulation of its fan law. For it can be intuitively seen, though not demonstrated, that if the point of origin of any fan is to be its total (ideal) dictionary-entry, and if that is to be thought of as its total meaning, then we can say that the meaning of any use of the fan which forms a point in the row of points of the fan, is included in the total meaning of the fan, and we can formalise this determination-relation as " \geq ". We can further say that the total dictionary-entry of the fan will have the form: "the use x^1 , and/or the use x^2 , and/or the use x^3 , up to x^n ".*

It now looks as though we can make two additions to the fan law; one defining the determination-relation of any

* Note for Linguists: Here I do not wish to go behind the actual texts of dictionary-entries, replacing them as they actually are by something which I, as a logician, am asserting that they ought to be. I am merely wishing to call attention to the fact that whereas scripted dictionary uses (e.g. in the O.E.D.) tend to be joined by commas or semi-colons, in colloquial dictionary-entries, (e.g. "The word 'plant', in Italian, can mean 'plant' or 'trick' or 'plan'") the connective word, joining the list of uses, will sometimes be 'and' and sometimes 'or': i.e. it can therefore compactly be referred to as "and/or".

fan as the inclusion-relation, and the other defining the hinge of any fan as the join (in the Boolean sense) of the points in the row of points in the fan. Moreover, having made these two additions to the fan law, it looks as though a mathematically reasonable state of affairs were beginning to set in; as though, for instance, we were getting into a situation in which we could ask, "Is a fan an oriented partially-ordered set?"

A moment's reflection, however, will suffice to show us that we cannot make these two additions to the fan law; because we cannot ever exemplify them. For the inclusion-relation, the partial-ordering relation, is defined as being reflexive, anti-symmetric, and transitive = that is to say, it obeys the three axioms: "For any x , $x \geq x$ " (the reflexive axiom); "For any x and y , $(x \geq y) \cdot (y \geq x) \supset (x = y)$ " (the axiom of anti-symmetry); and, "For any x , y and z , $(x \geq y) \cdot (y \geq z) \supset (x \geq z)$ ", (the axiom of transitivity). Moreover, the Boolean join relation might be defined as "Granted a partially-ordered set in which every two elements, x and y , have a least upper bound, P , and a greatest lower bound, Q , we can define P as ' $x \cup y$ ' (and Q as ' $x \cap y$ ').". But we cannot define either of these relations: not because we have no " \geq " or " \cup ", but because we haven't any "xs" or "ys". For the first clause in the fan law of any fan is, "A fan is a formal construction such that i) it has A sign (i.e. one single sign) which we will call the sign of the fan". We have further given an example of a fan, in Fig. I, a fan the dictionary-entry of which (that is the hinge of which) and all the uses of which (that is, all points on the row of points of which) have the sign . For  let us now substitute x. By defining the determination-relation of the fan as \geq we can now assert, that, in any fan, $x \geq x$ (that is, the meaning of the total dictionary-entry  includes the meaning of any separate use of ): that is, we can assert the reflexive axiom. But we cannot assert the anti-symmetric axiom, or that of transitivity, since the assertion of these requires, in the first case, two, and in the second case, three, signs (not uses, but signs), and by our definitions of fan and of sign, given in the fan law, we have, by definition, ourselves made impossible for any more signs than one ever to occur in the

fan. Of course, we can substitute the \underline{x} s and \underline{y} s for the uses, not the signs; we can say, "let \underline{x} stand for "plant used as plant", \underline{y} for "plant used as plan", \underline{z} for "plant used as trick?" But then the x s and y s, together with the uses which they symbolise, become part of the context law and cease to be part of the fan law; for if they become part of the fan law, firstly there would be no way of telling them apart (i.e. 'x' could certainly well be used for 'y')» and 'x' for 'z', in which case every formula would reduce to " $x \geq x$ "; and secondly, on the definitions, they would immediately become signs for other fans; and we are so far studying the fan in isolation; we have no other fans.

Now, given this difficulty of the looseness of fit between sign and signified, and which I believe, rightly or wrongly, to be the same difficulty as that brought forward by Bar-Hillel at I.C.S.I.)* what are we to do?

Put generally, my suggestion is that we must provide for every fan not only a fan law, but also a context law. Not just a set of contexts, but a context law.

In order to make clearer what I mean, I will do what I can to specify this context law. From this specification, it will become clear that the point of providing, for any fan or for any language of fans, a context law, as well as a fan law, is that only by interrelating what we know about the fan from the fan law with what we know about it from the context law, can we put ourselves in a position to increase our mathematical knowledge of the situation.

* No point brought up by Bar-Hillel has made him more unpopular than this one has: and there is none, logically speaking, on which he is more wholly right. On this, though, as on our capacity to recognise context when we ought not to be able to, logic doesn't have control of the whole story.

CONTEXT LAW OF ANY FAN.

1. Let the dictionary-entry of any fan, which is formally indicated by the hinge of the fan be....(here insert what the actual dictionary entry of the fan in question is).
2. Let the logical form of the dictionary-entry of any fan be presumed to be: (first entry) and/or (second entry) and/or (third entry)...up to n entries. Let these be formalised as (first entry) \cup (second entry) \cup ("third entry) \cup (nth entry).
3. Let the relation between the dictionary entry of any fan and its set of uses be that of inclusion of meaning, formalised as (dictionary entry) \geq (use in question).
4. Let the contextual meanings, or uses of any fan, which are formally indicated by the row of points of the fan be... (here insert for any fan, or set of fans, the actual set of uses of those fans) reading from left to right along the row.*

It is now clear what we want from our contexts (and that it is not the same as what the linguist wants of his collocations, or the literary critic by his contexts). What we want is a specification of context which will fit into item 4 of the Context Law of any Fan, (given above); that is, a specification such that any inferences which can be made, in the case of any fan, as the result of logical information obtainable by comparing or otherwise analysing the set of uses of the fan, will be thenceforth straightforwardly usable in the system⁺. And since, in abstract thought, to know clearly enough what you want is almost certainly to put yourself in a position to get it, we can

* The fact that the set of uses of any fan is ordered from left to right along the row of points, that is, in the same direction as the flow of time, gives the dictionary-maker the right logically to order the set of uses of any fan as he wishes. In most dictionaries, this logical order is given as the historical order of the development of the uses of the word in the language, but it is not clear that the two are always the same.

+ For example, suppose that of the set of two uses of a particular fan with a dictionary-entry "Cleavage", one use means "to stick together", ("she, cleaving only to him..") and the other use means "to split apart" ("split the stone and thou wilt find me, cleave the wood and I am there..") (This example is from Waismann). We can then say, from our knowledge of the context law of this fan, that one of these uses is complementary to the other; and if it has already been established that the mathematical system constituted by the fan was such that it could make sense to say of it that it was a complemented system, then the actual property of complementation could be added to it, from information derived from the context law.

now exemplify various well-known ways which the human race has dreamed up for making specifications of word-contexts, take the one we want, insert it into the theory, and then (and only then), having done all this, ask ourselves, "What is the empirical foundation for all this?" Thus an estimate of the empirical value of the notion as we have created it comes at the end, not at the beginning, of our exposition of context.

If I am right in thinking that the basic human language-making action consists in dreaming up fans; (that is, in first evolving logically primitive, i.e. general and indeterminate, language-symbols, and then, in explanatory talk, specifying for them more and more contexts); it will follow that the various devices for specifying word-use in any language, will be the logically primary devices of the language. And so, they are; the pointing gesture, the logical proper name ("Here!" "Now!" "This!") the defining phrase, all these are logically far more basic than case-systems or sentence-connectives. In short, in asking for the kind of context-specifications which I am looking for, what I am after is the most logically primitive form of definition.

This can be obtained instantly the moment it is seen that the basic characteristic of definitions is that they don't define. They distinguish, just as a pointing gesture does, but they don't distil. Except possibly in mathematics, which we are not now talking about, you can never go away hugging your definition to your breast, and saying, "Ah, now I've got THE meaning of that word!"

As soon as one has thought this thought, one achieves liberation, in that one ceases to look for merely one kind of definition. One lifts one's eyes, and says, "Well, how do people distinguish word-uses from one another?"

1. They do it by gestures, especially when they don't know the language. (We won't go further into this, now),
2. They do it by explanatory phrases, "'Father' usually means 'male parent'. But it doesn't always. 'Father' can mean any venerable person. The Catholics use it as a name for priests.", and so on.

3. They do it by actually showing the word in the use which they want to distinguish. 'Rich' means 'humorous'; have you never heard the phrase, "That's rich"?

It is upon this fact, - namely, that exhibiting a word in collocation is one well used type of context specification, that the scientific linguist bases his hope of getting meaning distinction from texts. Well, he may; and this would give us at once an empirical definition of context; but he hasn't yet.

The kind of difficulty I believe him to be up against can be exemplified by the way in which I learnt the meaning of "That's rich!" I learnt it when a sudden spasm of laughter at a joke suddenly convulsed me; and someone else, who was also laughing, said "That's rich". In other words, I connected the phrase, "That's rich" with a kinaesthetic sensation, that is, with an extra-linguistic context, not an intra-linguistic one. The fact that "rich" occurs in this sense, often in the collocation "That's..." was irrelevant, and is to my distinguishing this meaning of "rich".

4. They do it by compiling lists of synonyms: "Father, male parent, male ancestor".

This is a special form of procedure 2, and in my view, it is a perfectly valid convention of definition. Why should you not just group overlapping word-uses, and then say no more, instead of giving each a lengthy explanation.

5. They do it by juxtaposing analogous sentences. I have treated of this in my companion-paper in this volume. It is the method currently used by what is currently called derisively "Oxford philosophy"; that is, by the current school of philosophers of ordinary language.

If we now recall the whole argument of Section I, it will be clear that the kind of specification which will give our fan, or any set of fans, a context law, is the synonym-compiling device given above under 4). If the synonyms in such groupings were complete synonyms, the device would be no use to us; but they are not. They are dis-

tinguished one from another, by being e.g. more colloquial, by being e.g. pejorative or approbative, or more intensified versions of one another; and the groupings are distinguished by sentential function. In short, the synonyms in synonym-groupings are compared to one another and distinguished from one another in terms of specifications by heads, syntax-markers, archeheads...

To sum up: whether you decide that context, in this sense, is an empirical notion, will depend firstly, on whether you think that the five forms of definition which are given above are logically equivalent; and secondly, whether you think that any one, (say, 3, or even 5) could be explored by detailed research-methods to throw light upon 4*. If you think that either could, you will be empirically satisfied; and even if you do not think this, you need not be ultimately dissatisfied, if a context-system, successfully built of language-fans achieves mechanical abstracting or M.T. For basically, a word-use in context is something which you "see"...

* In the C.L.R.U. Library Scheme, Mark II, now being redesigned, a mechanical procedure for computing similarity of word-uses from term-abstracts, and designed by R.M. Needham, is the chief feature of the redesigned scheme. Each time two words are found together in a term-abstract, they score 1 for similarity; pairs of similarities are thus the units of the scheme. Term-abstracts can be non-contentiously compiled either by human-beings or mechanically; they are the most frequently-occurring words in any document.

Note: In order that this paper shall be included in this volume, the theoretic descriptions of Head, Archehead, Row and List have been included here in summary form.)

HEADS.

a. It should be possible, by taking the notion of Fans, to construct a generalised and weaker version of Brouwer's calculus of Fans.

If this could be done, then Brouwer's Fan Theorem,⁽¹⁾ which in classical form is the stop-rule theorem in Koenig's Theory of Graphs,⁽²⁾ will provide a theoretic definition of Head.

(I say: "If this could be done": I cannot at present do it.)*

b. The question has to be discussed as to whether the totality of contexts in a language form a continuum, in view of the fact that the set of contexts of any word appear to form a discrete set. That is to say, if a word is being used in one way, it is not being used in another. The uses of a word do not "fade into" one another; new uses continually appear, but the set of them is discontinuous.

As against this, I can see no way of imagining the total set of concepts of a language (i.e. the set of the total possible continually-increasing dictionary-entries of all the words) except as a Brouwerian continuum.

* Instead of mapping on to the rational grid, you have to map on to a lattice. Then the proof must consist of saying: i) that a mathematical proof also can be mapped onto a lattice, which I believe it can; ii) that this lattice is a proper sub-system of the total system; iii) that since a proof has a determinate end-point, reached in a finite number of steps, so does the system; iv) that this end-point of the system is the Head.

1. L.E.J. Brouwer, "Points and Spaces", Canadian Journal of Mathematics, 1954. N.B. Not as commented on by A. Heyting, "Intuitionism".

2. This equivalence has been shown independently by S.C. Kleene and R.B. Braithwaite.

Because of this, my present view is: make a continuum, (Brouwer's is the only true continuum) and then use the context-law to wrinkle it afterwards.

c. The question has to be discussed about the empirical status of heads. They contrast with contexts; contexts, or word-uses, look very empirical until they are subjected to analysis, when it turns out that you have to "see" them. Heads, on the other hand, gain empirical solidity the more the notion of extra-linguistic context is analysed, and the more thought is given to the practical necessity of accounting for human communication. (Roughly: something must be simple and finite, somewhere.)

Probably, perversely, I have hopes of confirmation for this part of the theory coming from research in cerebro-physiology.

A paragraph describing the way in which heads introduce finiteness into the system is given below. (This paragraph is taken from a paper which I have written but is not yet published.)

"Philosophically, it comes to this: the fundamental hypothesis about human communication which lies behind any kind of thesaurus-making is that, although the set of possible uses of words in a language is infinite, the number of primary extra-linguistic situations which we can distinguish sufficiently to talk to one another in terms of combinations of them, is finite. Given the developing complexity of the known universe, it might be the case that we refer to a fresh extra-linguistic situation every time we create a new use of a word. In fact we do not; we pile up synonyms, to rerefer, from various and differing new aspects, to the stock of basic extra-linguistic situations which we already have. It takes a noticeable new development of human activity (e.g. air travel) to establish so many new strings of synonyms in the language that the thesaurus, Aerial Motion may conveniently be promoted from being a subhead of Travel to being a new head in its own right; and even then, if inconvenient, the promotion need not be made.

"The primary noticed universe remains more stable than do continually developing sets of uses of words; in fact, all that ever seems to take place in it, in the last analysis, is a reorientation of emphasis, since the number of heads in any known thesaurus never increases beyond a very limited extent.

"The importance of this fact for Machine Translation, is obvious. If the hypothesis is right, communication and translation alike depend on the fact that two people and two cultures, however much they differ, can share a common stock of extra-linguistic contexts. When they cannot come to share such a stock, communication and translation alike break down. Imagine two cultures, one, say, human, one termite. The members of the first of these sleep, and also dream, every night; the members of the second do not know what sleep is. As between these two cultures, communication on the subject of sleeping and dreaming would be impossible until acquired knowledge of sleeping and dreaming by members of the second culture sufficed to establish it."

3. Archeheads.

(Examples of archeheads, from Richens' *Nude*, are given in Section IV.)

The problem of theoretically describing an archehead involves bringing up the difficult notion of the meaning-line.

a. The problem of the meaning-line.

It is found in practice, that when points in the thesaurus-lattice are very near the top, they become so general that, by meaning practically everything, they cease to mean anything. Such points will be defined as being "above the meaning-line". In practice, we count them, or call them by letters, or by girls' names, ("Elsie", "Gerite", "Daisy"). Each of these devices, (see Section I, above) is strictly speaking, logically illegitimate, in that it ascribes to such points a type of particularity which they haven't got. It isn't that they mean nothing: it is that they mean too much. They are, in the logical empiricist sense of the words, metaphysical.

b. Archeheads must be just below the meaning-line.

They aren't words which could exist in any language. But they must be sufficiently like words which can be handled in any language to enable them themselves to be handled. TRUE! must be like true; or at least, TRUE! must be more like true than it is like please.

Until lately we were so impressed by this difficulty that we assumed that it was impossible, in practice, to name or handle archeheads. Constructing Richens' Nude has convinced us that this can be done.

R. H. Richens is thus the discoverer of archeheads, not as theoretic entities (they are in Roget's chapter of contents) but as usable things.

c. Archeheads, as has been shown by tests on Nude, have an extremely practical property: they intersect, when the thesaurus algorithm is applied to them, at just those points where the thesaurus itself lets you down:

e.g. change/where | in(pray:where:part) - CHURCH

this is "to go to church", in Nude. Notice that the archehead WHERE! is here in common between both entries: although you would never persuade a thesaurus-maker to include "church" in a list of places to which people go.

e.g. (cf. Bar-Hillel)

in | (man/use)/(in:thing) - INKSTAND

"in the inkstand"

Notice that the archehead IN! is in common between the two entries, although no thesaurus-maker would intuitively think of "inkstand" as an in-thing unless something had brought the fact that it was to his notice.

These intersections, of course, are caused to occur by the fact that, if you have only 48 archehead-elements to choose from in defining something, the chances go up that descriptions will overlap.

In other words, the fewer the heads, the smaller the resolving-power of any thesaurus; and the smaller the resolving-power of any thesaurus, the greater the intersecting power of the thesaurus. In order to combine a high resolving-power and a high intersecting-power, the thesaurus should contain a large number of heads, to secure the first, and, including them, a large number of archeheads, to secure the second.

Thus, a thesaurus of 48 heads, which is what Nude can be taken as being if you ignore the sentential connectives, has a very high intersecting-power indeed.

4. Rows.

The problem of making a theoretic description of a row is that this involves making a theoretic description also both of a word, and also of a language.

(Actual examples of rows are given in Section IV.)

For a) the rows of a thesaurus consist of words, (but these words can be of any length),
b) the totality of rows of the thesaurus (empirically speaking) constitutes the language.
And how do we distinguish here "languages" from "language"?

i. Words.

The great difficulty of defining a "word" has been discussed by me some years ago in a publication*. I pointed out there that nobody has, in fact, tackled the problem of defining the notion of a "word" in an intellectually satisfactory manner. Philosophers regard it as being purely a grammatical concept. Traditional grammarians are leaning on what they believe to be the insights of philosophers; modern linguistics professes not to be interested, for it claims that the "word" is in no sense a fundamental notion.

* "Words", Proceedings of the Aristotelian Society, April, 1954-55.

So the difficulty is there, in any case. If the thesaurus is to be interlingual, there is no one length for "word". As so often, the difficulty of operating within one language mirrors the difficulty of operating between various languages.

One's first impulse is to say, "Let a word be any stretch of language, short or long, which, in practice, serves to distinguish a point on Rank of the thesaurus-lattice."

But this definition is circular. First, we define the points on Rank V of the thesaurus-lattice as being those separable words the contexts of which can be mapped on to the points of Rank 4: then we define the words which go on a thesaurus-lattice as language-stretches which map on to the points of Rank 4 of the thesaurus-lattice.

I do not see the way out of this difficulty.*

ii. Language.

a. Language is an abstraction. All logicians know this; but they behave as though the "fit" between the abstraction "Language" and any language is so close that the fact that "language" is an abstraction doesn't matter.

Nothing could be further from the truth. The proposition "Language exists" is a theoretic one. It is rather like, "Matter exists", or "God exists", or still more, "The Universe, considered as a whole, exists".

What is needed is a theoretic definition of "a language".

b. What we know about a language, according to the theory, is that it is a sub-lattice of the total language-lattice. The archeheads, the syntax-markers, the heads of any given language will be a different subset of the total set, but each will be a subset of the total set.

Footnote to p. 38:

* It should be possible to find a way, by using the Fan-Calculus. For the Context Law has an analogy with Brouwer's Complementary Law; and it should be possible to use this analogy (somehow) to construct a context-law-derived entity, in terms of which we could then theoretically define word, and so escape the circularity given above. This is not the part of the theory which, in the final statement, most needs to be complete and right; and this is just the part on which I have no light at all.

Yes, but suppose what is really different as between language and language (considering now "a language" as well as "Language" as something which is given in terms of the theory) is not that it is made up from a different set of archeheads, markers, heads, but that it is made up of these in different combinations? This would mean that every language was a different lattice, not a sub-lattice of a central total language lattice⁽¹⁾, and that every single language-lattice had different rows. The semantic, grammatical and syntactic devices used by any given language would then be imagined as being alike, distinguishable and specifiable in terms of combinations of a set of initially very weak semantic components. These components would be very like indeed to the weak semantic components which linguists at present use to distinguish components of a system.

It has frequently been claimed by linguists, particularly those of the American "Structuralist" school, that their subject is a science, based on purely empirical foundations; some have even gone so far as to describe it as a kind of mathematics. However, it is impossible to relate the abstract systems linguists create to any particular linguistic situation without reference to immediate and undisguised concepts. As Kay has said, in the companion paper in this volume, the moment one asks the most fundamental question of all, "What is being said here?" we must find other apparatus than linguistics provides. Thus, it is that when Harold Whitehall writes on Linguistics as applied to the particular case of the English language⁽³⁾ semantic categories, heads, descriptors - call them what

1. They will all be sublattices of the lattice of all possible combinations, but this lattice is both almost unconcernedly large and also empirically irrelevant.

2. Numerous reference to claims of this sort are given in "The Relevance of Linguistics to Machine Translation" by M. Kay. In particular, see Martin Joos.

3. "The Structural Essentials of English", Harcourt, Brace and Co., New York, 1951.

you will - immediately begin to play a leading part. One of the great merits of this book, in my view, is that no apology is made for the introduction of these semantic categories; they do not have to be introduced furtively under the guise of mnemonics for classes established in a more respectable way. The following is an actual table from Whitehall's book:

Fig. 4
THE SYSTEM OF PREPOSITIONS⁽¹⁾

RELATION	Simple		Complex	Double	Group
	Primary	Trans-ferred			
1. Location	at	down	aboard, above,	inside, outside	in back of
	by	from	across, after,	through-out,	in front of
	in	off	against, amid,	toward(s),	inside of
	on	out	before, beneath	underneath,	on board(of)
		through	beyond, near,	upon, within	on either
		up	beside, between	without; down at	side (of)
			next, over, past	at, by, in, on;	on top of
			under	out at, by, in,	outside of
				on; up at, by,	
				<u>in, on.</u>	
2. Direction					
	down	at	aboard, about,	inside, outside	in back of
	from	by	across, after,	toward(s); under-	in front
	off	in	against, among	neath; into, onto,	inside of
	out	on	around, between	down to, from	on top of
	through		beyond, over,	off to, from;	on board(of)
	to		under	out to, of, from;	on either
	up			up to, from; near	side(of)
				to, next to; over	outside (of)
				to; to within,	
				from among	

Similar tables are used in Viggo Brøndal's "Theorie des Prepositions"⁽²⁾

So, looking at this fundamental feature of linguistics from a theoretic and thesaurus-maker's point of view, we see that Einar

1. Whitehall, op. cit., p. 72.
2. Copenhagen 1950.

Haugen may have been onto a more important point than he realised when he said

"It is curious to see how those who eliminate meaning have brought it back under the covert guise of distribution."⁽¹⁾

The discipline which we are here imposing on the linguist is that we will not allow him a fresh set of concepts for each system. His semantic concepts must form a single finite system; and with combinations of them he must make all the distinctions which may turn out to be required within the language.

Now, if word and language can be theoretically defined. as I have desired to define them, but failed to define them, above, then we can say that a row is a set of overlapping contexts of words in any language, this set being distinguished from all other sets in terms of heads, markers, and archeheads, but the members of the set only being distinguished from one another by means of archeheads.

To go back to the question of each language being a separate lattice, instead of each being a sub-lattice of a total language-lattice: this does not seem to me to matter as long as the lattice-transformation which would turn any language-lattice into any other is finite and mathematically knowable.

iii. The row is also an empirical unit in a thesaurus. You test for rows, as a way of testing Nude and Lattite. If a thesaurus or interlingua, when used on any language, produces, when tested, natural-sounding rows and lists which really occur as lists in that language, then the thesaurus or interlingua has an empirical basis for that language. If the test produces arbitrary collections of words, then the thesaurus is arbitrary.⁽²⁾

The empirical question as to whether in practice rows can be found which are interlingual, is discussed to the extent to which I am able to discuss it, in Section IV.

1. "Directions in Modern Linguistics", Lg 27 (1951). Presidential Address to the; Linguistic Society, Chicago, 1950.

2. This test works, too. You know at once when you see the set of cards, whether it is trying to be a list or a row, or whether it is arbitrary.

5. List-Numbers.

a. Lists are sets of mutually exclusive contexts.
e.g. spade, hammer.

If he hit her with a spade, he didn't hit her with a hammer. In the sentence, "He hit her with a , " either "spade" or "hammer" can be used to fill the gap, but not both (Contrast the sentence, "He was a coward, a craven, a poltroon".)

If one sentence mentions 2 members of a list, then the two members must be joined by at least "and". "He was carrying both a spade and also a hammer."

You can, of course, replace the commas by "ands" in "He was a coward, a craven a poltroon". But the "ands" won't mean the same thing here. The list-joining "and" is logically a true Boolean join, "and/or"; the synonym-joining "and" is a logical hyphen, a meet. You might say, "He was a coward-craven-poltroon",

b. Theoretic definition: a list-number is a head in the thesaurus with only one term in it; that is, with only one context, or word-use in it.

Thus, the sub-thesaurus consisting of the members of a list is, and always will be, a spindle. The occurrence of a list-number in a thesaurus-using translation programme, is a warning that the limit of the resolving-power of the thesaurus has been reached.

2. Algorithm for the translation of list-numbers.

Take the thesaurus dictionary-entry for "carrot".

Take also the dictionary-entry for "parsnips".

These two dictionary-entries are saved from being identical by the fact that you can dangle a political carrot in front of someone; and that "Hard words butter no parsnips". So the two words can be distinguished from one another, in the thesaurus, by the fact that they do not have identical dictionary-entries. But the two contexts cannot be distinguished from one another when both of them occur in the same row of head VEGETABLE. Suppose we try to translate the following sentence, "He was digging up a carrot in his garden", then the translation-algorithm will pro-

duce the whole list of vegetables.

The only solution is to add to the dictionary-entry of carrot and parsnip a list-number which is attached to a definite head of the thesaurus (say, VEGETABLE) but does not have to intersect in the intersection procedure. Thus, carrot, as well as having a political head in its dictionary entry, will also have VEGETABLE, (139). And parsnip, as well as having a civility and soft-spokenness head in its dictionary-entry, will also have VEGETABLE, (141). As soon as the translation-algorithm gives VEGETABLE as the context, the machine picks up the list-numbers. It then brings down the list given under VEGETABLE, and brings down the one-one translation of carrot into the output language, of carrot given under (141). In other words, a thesaurus list is a multi-lingual one-one micro-glossary (no alternative variants for any list-word being given) in which the different members of the list have different numbers. But the micro-glossary itself must be attached to a given head; because only when it is known that that head gives the context which is being referred to in the input text, as it is known also that the words in the micro-glossary will be unambiguous. "Mass" can mean "religious service" as in "Black Mass"; "charge" can mean "accusation", or "cavalry-charge". Only when it is known that both are being used in the context of physics can they be translated micro-glossarywise, by using their list-numbers.

d. Theoretic problems which arise in connection with list numbers.

It might be thought that the theoretic problems of list-numbers would be easy. Actually, they are, on the contrary, very difficult; and the philosophy of lists is still most imperfectly understood.

Certain things are known:

i. No head must contain more than one list; otherwise the procedure* will not tell you which list to use. If you

* The difficulty is a coding one; methods may perhaps be found to associate a list with a combination of heads.

want more lists, you must have more heads,
ii. One word, however, can figure in several lists.
iii. The list-procedure, unlike the translation-algorithm,
gives a single translation.

But none of us really knows how to compile a list when
it is safe to have a list, when not; and what is the
principle uniting the words in a micro-glossary.

If the arguments of the above sections had been fully
filled out, and if all the difficulties arising
from them had been adequately encountered, this would
be the end of the theoretic part of this paper.

In the section immediately following this paper, and
the one after, the problems brought up for discussion
are much more empirical problems.

III. KINDS OF THESAURUS.

1. Bar-Hillel, and other critics, have asserted that the C.L.R.U. uses the word Thesaurus in a variety of different senses, thus causing confusion. This criticism must be admitted as correct. It can also be correctly replied, that these senses are cognate; and that different senses of "thesaurus" are being used, because C.L.R.U. is experimenting with different kinds of thesauruses. The purpose of this section is to enumerate and describe the kinds of thesaurus, so that the difficulty caused by past inexplicitness may be overcome.

All the kinds of thesaurus, which are used in the Unit, can be taken as being partial versions of the total thesaurus model defined in Section I above. This provides the unifying theoretic idea against which the various examples of partial thesauruses should be examined.

The senses in which "thesaurus" has been used, apart from the total sense of Section I are:

- i. A natural thesaurus - e.g. Roget.
- ii. A term thesaurus - e.g. that associated with the C.L.R.U. Library Scheme.
- iii. An interlingua - e.g. Richens' interlingua.

2. The natural thesaurus. For most English-speaking people, this is exemplified by Roget's Thesaurus of English Words and Phrases (London, 1852 and later). In this document, words are grouped into 1,000 heads or notional families; words often coming into more than one head. An index at the back contains an alphabetical list of words with the numbers of the heads in which they come. There are, however, a number of other such documents:

- a. "Copies" of Roget in some 6 other languages (See Sect. IV).
- b. Synonym dictionaries. These are alphabetical lists of words with a few synonyms or antonyms attached. Heads could be compiled from these, but prove inadequate in practice.
- c. Ancient thesauruses. Groupings in language (Chinese, Sanskrit, Sumerian), where alphabetical dictionaries are ruled out by the nature of the script, have been found to have thesauric properties, though they may

be sometimes overlaid by the groupings round graphically similar characters. The best known of these is the Shuo Wen ancient Chinese radical dictionary.

While natural thesauruses have the advantage for experimental purposes of actually existing in literary, or even in punched-card form, (for which reason all C.L.R.U. thesauric translation tests have been made on them), they suffer from serious drawbacks imposed in part by the necessities of practical publishing. These drawbacks may be listed as follows:

- a. The indices are very incomplete. It seems that publishers insert only some 25% of the available references to the main texts since, if they insert more, the resulting volume is too heavy to publish. As for testing and mechanisation purposes, by far the most convenient way of using the thesaurus is to compile it from the index, this is very considerable research defect.
- b. Since the main purpose of thesauruses published in book form is to improve the reader's knowledge of words, they tend to leave out everyday and ordinary words, and to insert bizarre and peculiar words which will give the user the feeling that his word-power is being increased. For translation purposes, the opposite is what is required.
- c. In Roget, the "cross-references" from one head to another are very incomplete and unsystematic. Their insertion causes an even greater inadequacy of the index; their omission, an even greater dearth of ordinary words in the heads.
- d. The heads themselves are classified, in the chapter of contents, by a single hierarchy, in tree form; whereas what is required is a multiple hierarchy of archeheads. The cross-references between heads provide the rudiments of an alternative classification; but this is too incomplete to be much use.

All these deficiencies may be discovered by simply opening and reading an ordinary Roget. More recondite characteristics of the existing document were brought to light by tests of various kinds.

e. The cross-references from head to head tend to be symmetrical; that is, a head which has a great many cross-references from it is likely to have a great many cross-references to it. (C.Wordley's punched-card tests.)

f. The intersection procedure, as in "Agricola.." (Appendix III to Bastin & Needham) failed to work even when reasonably predictable common contexts were present, in an attempted translation from English to English. This was almost certainly because the common possibilities of word combination in the language are not in it. (See Section II, on Archeheads).

g. The thesaurus conceived as a mathematical system was exceedingly redundant, and when this redundancy was investigated further, it was found that this was because of the presence of a large unordered profactor in the lattice containing the thesaurus. (Parker-Rhodes & Needham, "Encoding Roget's Thesaurus," C.L.R.U. workpaper; cf. also the essay in this collection). This was tantamount to saying that the thesaurus at present existing had a great deal less usable structure than would at first sight appear.

h. Some of the heads can be shown by tests to be arbitrary. Most of the arbitrary heads are artificial contraries of genuine heads. As a result of all these characteristics, although the idea of a thesaurus is sometimes most conveniently defined by displaying Roget as a particular example, it becomes clear that existing thesauruses are very unsuitable for M.T. work. However, it is possible from the defect above to obtain a fairly precise idea of the changes that are necessary to make a usable thesaurus for mathematical treatment. It is likely that for some time to come experiments will make use of the natural thesauruses with changes made to remedy particular defects, rather than with an entirely new thesaurus which would require a major effort for skilled lexicography which will in turn require a considerable time to carry through.

2. The term thesaurus. The term thesaurus is exemplified by the thesaurus used for the C.L.R.U. Information Retrieval System (Joyce and Needham, Amer. Doc. 1958). It was invented to deal with a situation where a large number of new technical terms had to be handled which were not to be found in any existing thesaurus (or, for that matter, any dictionary). Also, it was required for reasons set forth in Joyce and Needham (loc. cit) that all terms should be retained as individuals as well as being incorporated in heads, while nonetheless, all reasonable heads should be used. The structure thus set up is a very detailed one, with a large number of levels. There is no formal distinction between heads and terms, and the thesaurus (which is sufficiently small, can actually be drawn on a rather large piece of paper) appears a multiple hierarchy of points representing words, The point representing word A appears above a point representing word B, if the uses of A are a set of contexts including those of the word B. In many parts of the system, this corresponds to a straightforward subject classification, which is clearly a subcase of the whole. It will be seen that since each word is treated entirely individually, the degree of detail of the system is rather greater than that of natural thesauruses; the term thesaurus can cater for relations of considerable complexity between words which would simply fall under a head together in the natural thesaurus. A sample of the classification system is attached, consisting of the sublattice concerned with the request for documents on "Linguistic Analysis and M.T, analysis". Here all the terms used are fairly high up the hierarchy, and only a few of the 16 levels in the hierarchy are exemplified.

The operation of this kind of system is discussed in detail in Needham and Parker-Rhodes, (essay in this collection). There is some advantage, however, in here discussing it again, in order to consider the relation of the system to the other kinds of thesauruses.

Firstly, it is clear that the higher terms are functioning as something very like heads (or even archeheads), as well as functioning as words in their own right. It has appeared

(Miller, "Extension and Testing of C.L.R.U. Library System") that this phenomenon has in some cases seriously warped the lattice in the sense that a term high up (e.g. mathematics) carries so much weight by virtue of the many terms that it includes that it no longer functions efficiently, as the terms associated with its word (e.g. "mathematics") This defect may be corrected by using a device; however, it indicates that the treatment of all words and heads, pari passu may be incorrect.

Secondly, the system is excessively cumbersome through the great number of its terms; in an anxiety (Joyce & Needham) not to lose information from the system an uncomfortably large amount has been kept, much of which is unlikely to be required. Now this was an anxiety not to lose it by absorption of words into entirely intuitively based heads. The intuitively based heads are there, expressed by the inclusion system of the lattice; but the original and detailed information is there too.

It is at present intended to conduct experiments on the mechanical reconstruction of the retrieval thesaurus, which are expected to throw considerable light on the relations between the term thesaurus and the natural and total thesauruses, and also to throw more light on the structure of the latter. The basis of these experiments is the idea that words which can properly be amalgamated in a head should have the property of tending to occur together in documents; if the heads are built up on this principle the loss of information through replacing the word by the head will be minimised. This naturally gives rise to a measurement of the extent to which pairs of words tend to occur in the same documents, which will be called their similarity. In order that experiments may be made to see whether this line of thought is at all profitable, two things are necessary:

- i. An algorithm for calculating on some agreed basis in the data what the similarity of a pair of terms shall be,
- ii. An algorithm for finding, from the total set of terms, subsets which have the property that the similarity between their members are high compared with similarity between members and non-members.

Several algorithms of the type "i" are available. Probably the simplest is that described by Tanimoto (An Elementary Mathematical Theory of Classification and Prediction, IBM Corp.). This may be exactly described if the agreed basis for computation is the description of documents by their term abstracts. The search for an acceptable rigorous definition and consequent algorithm "ii" is being carried on by several workers under the name of research into The Theory of Clumps⁽¹⁾. This is not the place for an extensive discourse on the progress to date in this field; however, various attempts exist. It is shortly intended to carry out by means of a computer an exhaustive examination of a simple case to compare them. If the results of this are satisfactory, tests will be conducted on parts of the C.L.R.U. Library Scheme, the general principle being as follows: An already-existing classification of the terms will be used as a kind of "trial set" of heads. On the basis of similarities of terms computed on an increasing number of documents; these heads will be examined for satisfaction of the "clump criterion" (as the rigorised definition "ii" is called), and altered so that they satisfy it as far as possible. These altered heads will then be used for retrieval.

3. Interlinguas.

An interlingua means here:

- a. A thesaurus consisting solely of the archeheads of Section I.
- b. A thesaurus with a procedure for finding out syntactic structure.

If the syntactic structure procedure is regarded as something super-added to the thesaurus, R.H. Richens' Nude is an interlingua in the present sense. If the bonding⁽²⁾ be disregarded, the 48 elements seem very like archeheads, and

1. The term "clump" was invented by Dr. I. Good.

2. Nude is described below in Section IV.

would give rise to a lattice structure with much less resolution than a whole thesaurus, but with an additional intersecting power⁽¹⁾.

An Italian-Nude dictionary of some 7,000 chunks has been made at C.L.R.U., and various tests on it have been performed. Since, however, only a small part of the dictionary has been key-punched, the tests have had to be limited and particular ones, directed to examining the internal consistency of the Nude entries for Italian. Typically, a set of near-synonyms was found from an Italian synonym dictionary, and their Nude equivalents found. These would come from different parts of the dictionary, and were usually made by different people; the object of the exercise was to see whether the entries were widely divergent. While the tests sometimes brought out errors of considerable differences of interpretation, in general, support was given to the objective character of Nude as an interlingua. These tests are to be continued and the detailed results written up.

While Nude conforms to the definition of a partial thesaurus, it suffers from the drawback that it has so far proved impossible to attach a quantitative measure to the extent to which one Nude formula is like another. If all brackets and bonds are removed so that the measures used in the total thesaurus may be applied, the results are unsatisfactory since much of the character of a word resides in its bonding pattern. The discovery of a procedure for "inexact matching" as it is called is a matter for present research on Nude, and when some progress has been made in it, it will be possible to repeat in a more cogent manner the tests on near-synonyms described above.

On the other hand, - though this is not a thesauric property - the fact that every Nude formula has a unique, though simplified, sentential or phrase structure, is of the greatest

1. Cf. Section II, 3, above.

help when Nude is used for translation. This is a characteristic which every attempt is being made to simulate in the full thesaurus, by establishing convert ability between the Nude sentential signs and certain combinations of elements in Lattite. No tests, however, have been done on Lattite as yet; so Nude remains the Unit's M.T. interlingua. The following (see overleaf) Italian-English translation trial, done on a randomly chosen paragraph with a dry run, probably gives a fair idea of what its translating power is. It is hoped that in the not too far distant future to put Nude on a big machine, in which case, large-scale Italian-English output could be obtained.

4. From the above accounts, it will be clear that, though we are indeed at fault in having used "thesaurus" in our reports in different senses, yet these senses are more cognate than might at first sight appear.

SPECIMEN TRANSLATION

Italian -> Interlingua -> English

Input

Il colere della farina caratteristica cui nel commercio si attribuisce assai grande importanza, dipende essenzialmente dalle sostanze coloranti naturali presenti nella stessa farina. Però sul colore varie cause accessorie influiscono e soprattutto la presenza di sostanze scure estranee. La granularità stessa della farina ha un effetto sul colore, giacchè i grossi granuli proiettano un'ombra che da alla farina una sfumatura bluastra.

(Genetica Agraria 1946:
1:38)

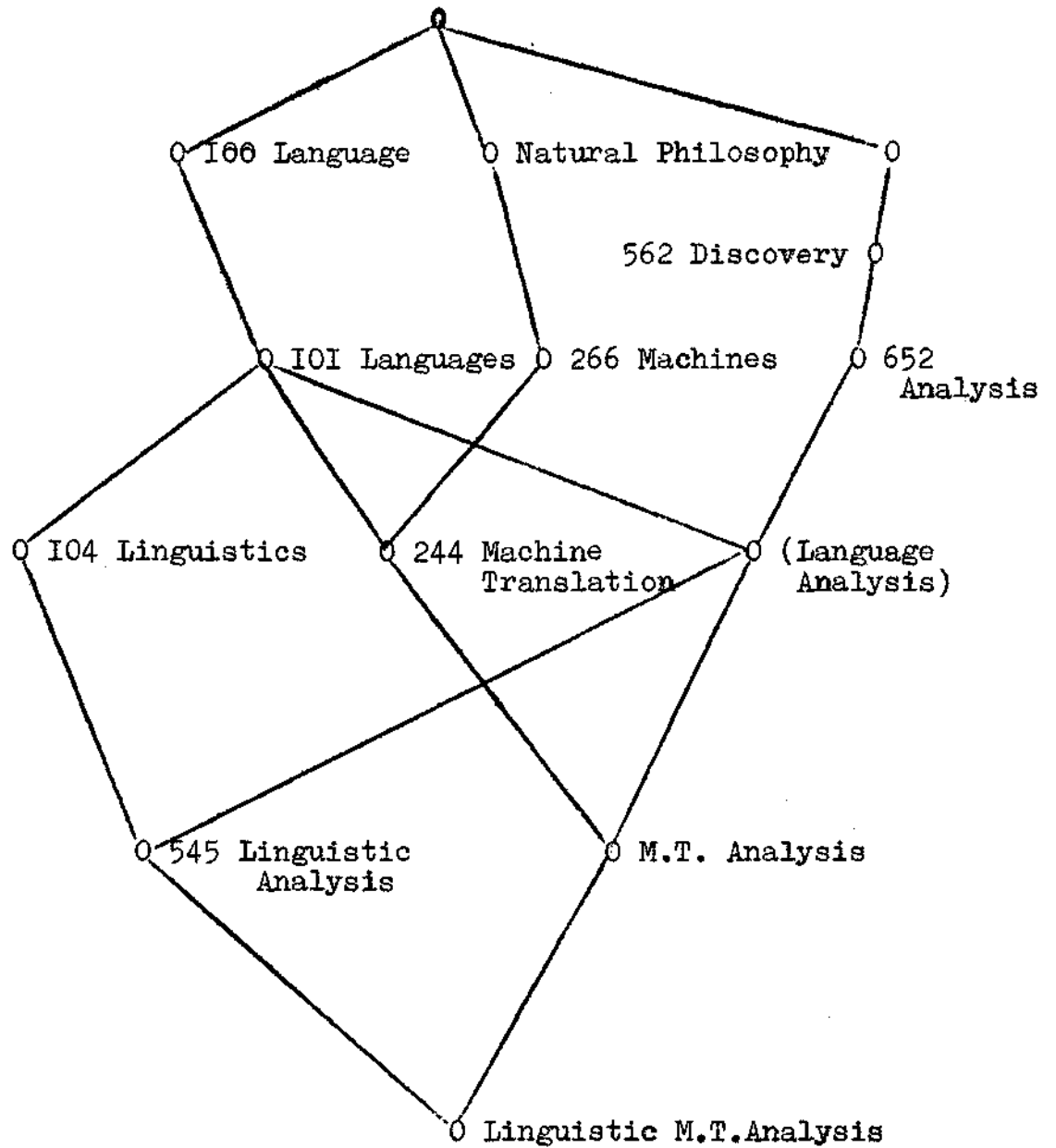
Output

The colour of the caratteristica flour of which very big importance is thought in connexion with commerce is condition-ed natur-ly by the color natural present substance in the same flour. But different accessor caus-es and especially presenc-e of dark estrane substanc-es influenc-e the colour. The same granul-ness of the flour has a effect in connexion with the colour because the big granul-s proiett a shad-e that giv-es the flour a blue-ish sfumatur.

NB. 1 - Words underlined did not occur in the dictionary used.

NB. 2 - In the above translation, caratteristica, which did not occur in the dictionary, was taken as an adjective. The correct interpretation is indicated by the comma, which precedes instead of following the word. Since commas are used so diversely, they have not been exploited in the present programme.

A Specimen part of the Library Retrieval Lattice.



Linguistic Analysis \cap Machine Translation \cap Analysis gives the following two papers:

- 1) Accession No. 72. Linguistic Analysis and M.T. Analysis.
By.P.Garvin.
- 2) Accession No. 352. C.L.R.U. Progress Report 1957.

TO WHAT EXTENT IS A THESAURUS INTERLINGUAL?

The extent to which any thesaurus is interlingual is, in practice, one of the most difficult possible questions to discuss. For two questions, which should be separate, always become inseparable. Firstly, "What would it be like for a Thesaurus to be, or not to be, interlingual?" And secondly, "So long as one and only one coded mathematical structure is used as the intermediate vehicle for translation, does it matter if it is, to a certain extent, arbitrary?"

1. The search for head-overlap between thesauruses in different languages.

The obvious first way to go about considering this double-headed question is to ask whether thesauruses exist for many different languages, and if they do, is there an overlap in their heads?

The immediately obtainable answer to this question is apparently most encouraging. Thesauruses with heads directly taken from Roget do exist in French, German, Hungarian, Swedish, Dutch, Spanish and Modern Greek.* This transference - and especially the transference into Hungarian, constitutes a high testimonial to the heads of Roget, - unless the heads in the first place could safely be arbitrary.

Now a procedure has been devised to test arbitrariness in heads. It was devised by Gilbert W.King, and was tried out on three subjects at IBM Research Mohansic Laboratory, York Town, New York in November, 1958. The heads selected were Cause, Choice and Judgement. The words from these heads were separately written on different slips of paper. 50% of them were left in piles to "define" the heads; the titles of the heads were not made known to the subjects. The other 50% of the words were shuffled and given to the subjects, who had to separate them back into their correct heads. All the three subjects proved able to do this with over 95% of accuracy. Moreover, they all titled the three heads correctly; and a misprint, "usual" for "casual", was without difficulty detected. Finally, a later attempt by one subject (the present author) to repeat the test, with the three heads Existence, Substantiality, and Intrinsicality failed; words like "real"

* This information, together with other information used in this section, comes from *Deutscher Wortschatz*, Franz Dornseiff.

"Hypostatic", "evident", "essential", "concrete", "matter of fact", "truth." etc., cannot be identified as belonging to any one, rather than any other, of the three. So it seems at first sight as though we have succeeded in contriving a simple and effective head-arbitrariness. It is all the more disconcerting, therefore, to find that it is the arbitrary heads, as well as the empirically founded ones as judged by this test, which are blithely transferred from Roget thesaurus to Roget thesaurus.

(And this immediately provokes the question, "Does it really matter? To what extent are these empirical entities at all?")

Let us next consider, in the search for head-overlap, the extant thesauruses which have not derived their heads from Roget. There is, for instance, "Der Deutscher Wortschatz nach Sachgruppen geordnet" by Franz Dornseiff, the "Dictionnaire Analogique" by M.C.Maquet, and various alphabetically ordered synonym dictionaries covering most of the European languages. These are encouraging to look at not only because there is a very considerable head-overlap between them and Roget; but also because the Roget heads which they have dropped are not the heads which it is likely that the test for genuineness, described above, would give as arbitrary.

There is less overlap, as one would expect, between heads of the ancient thesauruses and the modern ones. By the time one has documented oneself on the Amari Kosha and the Shuo Wen, however, and ignored the rumour that there is a Sumerian Thesaurus, and has asked why a Hieroglyphic Thesaurus has not been found, when they obviously had to have one, one is beginning to revive from one's first discouragement. One thing is clear; thesaurus-making is no evanescent or fugitive human impulse. It is, on the contrary, the logically basic principle of word-classification; the same principle as that which inspired the age-old idea of scripting a language by using pictographic or ideographic symbols. So, surely, something can be done to relate thesauri? Something which does not presuppose a complete cynicism as to the empirical foundation of the nature of the heads?

2. The procedure of comparing rows and lists.

In the special section on heads, in Section II, it was asserted that heads, by their nature, must represent frequently noticed extra-linguistic contexts. It follows from these facts that it is contexts, not facts, which are being classified.

and that the heads of a language are only the language users' frequently noticed set of extra linguistic contexts, not the total possible set of extra linguistic contexts. It follows that encyclopedic knowledge of all facts is not required by a thesaurus-maker, before he can assign word-uses to heads; but only a thorough knowledge of the contexts of the language.

This is all right in a theoretical exposition. As soon as one changes however, even in one's mind, from the very general word "context" to the more easily understandable word "situation", (thus replacing "extra-linguistic context" by "extra linguistic situation") then it becomes apparent that a sharper, smaller interlingual unit than that of a head is what is for practical purposes required. Consider, for instance, the comparable head-paragraphs, taken from an English, a French and a German Thesaurus respectively, and given below:

1. English: from Roget's Thesaurus: Head 739: Severity

N. Severity; strictness, formalism, harshness, etc. adj; vigour, stringency, austerity, inclemency, etc. 914a; arrogance etc, 885

arbitrary power; absolutism, despotism, dictatorship, autocracy, tyranny, domineering, oppression; assumption, usurpation; inquisition, reign of terror, martial law; iron heel, iron rule, iron hand, iron sway; tight grasp; brute force; coercion, etc 744; strong hand, tight hand.

2. French: Dictionnaire Analogique, edited by Maquet:

catchword Dur. (The catchwords are not numbered, being listed alphabetically.)

..Dur d'autorité Se faire craindre. Sévir, sévices; Maltraiter. Malmener. Rudoyer. Traiter de Turc à More. Parler en maître. Parler d'autorité. Ton impératif. Ne pas badiner. Montrer les dents. Cassant. Rembarrer. - Discipline. Main de fer. Inflexible. Rigide. Sévère. Strict. Tenace. Rigoureux. Exigeant. - Terrible. Tyrannique. Brutal. Despotique. - Rébarbatif. Pas commode. Grandeur. Menaçant Cerbère. Intimider.

3. German: Deutscher Wortschatz, edited by Wehle. Head 739 Strenge.

Härte. Unerbittlichkeit. Unerschütterlichkeit. Hartherzigkeit. Herzenshärte. Grausamkeit. Rücksichtslosigkeit. Gemeinheit. Unduldsamkeit. (Intoleranz). Rechthaberei. Herrenart. Schonungslosigkeit. Unnachsichtigkeit.

From a comparative inspection of these paragraphs two things become clear. Firstly, it is clear that the paragraphs are not

interlingual though the heads pretty exactly correspond; secondly, that the words could be rearranged so as to make the three paragraph-structures correspond a great deal more closely than they at present do. Moreover, there are two classificatory devices which could be employed here; firstly, that of getting the words of the same part of the speech, next one another, (and as has been already hinted in Section II, the relevant parts of speech in this particular case, are by no means as purely monolingual as they look); secondly, the further device of classifying words of the same part of speech by their "feel" (or aspect). "Traiter de Turc à More", for instance, and "rule with an iron hand" are both concrete images, both continuous processes, both pejoratives, both phrases indicating violence, both phrases describing a social habit of human beings. All these aspect-indicators are interlingual; there won't be a large class of word-uses in either language which have all of them: together with the head-reference, which in this case, is very highly interlingual, they may well jointly specify a single interlingual point. Nor is the comparative example which I have just given in any way exceptional; on the contrary, many paragraphs correspond more closely than these three.

Comparative perusal of thesauruses, then, shouts out for an interlingual way of defining paragraphs and aspects; and that without any concessions to preconceived theory. And if one is now determined not to be theoretical, the obvious method to start stream-lining paragraphs is in one's own language; and the way to do this, in each case, is to coin a descriptive phrase.

Below is an extract from an attempt by me to use this method to define a set of sub-paragraphs in Roget's Thesaurus which contain the word white. If it is desired to test my descriptions against other possible descriptions, all that is required is to cover up the right-hand column, in the table below, make your own set, uncover the column again, and compare.*

* It will be noticed that many of the row descriptions are verbal phrases, not noun phrases. The frequent use of these may be my personal idiosyncrasy; though the frequent appearance of such phrases in Nude entries also suggests otherwise; if the tendency to use verbal phrases for row definition is a natural one, then the criticism that a "thesaurus" is a system consisting only of nouns (G. King) is unfounded.

ROGET'S ROWS

whiteness

snow, paper, chalk, milk,
lily, ivory; white lead,
chinese white, white -
wash, whitening...

render white, blanch,
white-wash, silver, frost
white; milky; milk-white
snow-white, snowy, can-
did

white as a sheet; white
as the driven snow

vision, sight, optics,
eye-sight
visual organ, organ of
vision, eye

eye-ball, retina, pupil,
iris, cornea, white

abject fear, funk

white feather, faint-
heart, milk-sop, white
liver, cur, craven

faint-hearted, chicken-
hearted; yellow, white-
livered

etc.

DISCURSIVE DESCRIPTION OF ROW

people think of the abstract notion
of WHITENESS; a colour

white concrete objects, both solid
and liquid

the action of causing something
to become white.

people see objects having a white
appearance

concrete whiteness of colour being
used to symbolise mental states of
FEAR, INNOCENCE

the faculty of seeing

the part of the body with which a
man sees

list of parts of the eye

people exhibiting this

picturesque statements of the
appearance and physiology of people
exhibiting COWARDICE

people abusing their fellows in
concrete terms for exhibiting
COWARDICE

etc.

The question which this leads us to ask is two-fold: i)
could the descriptions in the right-hand column be expressed
in an arbitrarily-chosen language (I think they could). ii)
could a limited vocabulary be found for expressing them,
which itself could be translated into any language?

This limited vocabulary is what we hope Lattite is. Lattite
is the set of translatable mutually exclusive subsets of
syntax-markers and archeheads which is being used on the

thesaurus at present being multiply-punched on to cards. The reason why I am at present very coy about issuing definite lists of Lattite markers and archeheads is that until this thesaurus has been constructed and tested, it will be impossible to discover which of the Lattite terms turn out to define aspects, and which paragraphs and rows*. Instead of Lattite, therefore, I propose to discuss Nude, the simpler interlingua with 2 sentential connectives, 48 elements and two list-numbers, and nothing else at all.

(And again, the spectral question lurks in our minds: Suppose, whether using Lattite, or using Nude, different compilers give wholly different descriptions of the content of a row; either because they mistranslate some term of Lattite when operating Lattite in their own language, or because they 'see' the content of a row in a way differently from that in which other compilers 'see' it. Suppose this happens. Does it matter? Surely it does.)

We begin to suspect that, for translation purposes indeed it does matter. Indeed this question comes up so acutely in the case of Nude that it cannot be further deferred. Let us turn, then to the relevant features of this language,

3. The assumption which is being used for research purposes in testing how far any theoretically interlingual unit is, for practical purposes, interlingual, is the following: that divergences in assigning archeheads, syntax-markers, and heads, and which occur when different dictionary-makers assign interlingual specifications to words in their own language, sufficiently mirror the divergences in assigning them made by dictionary-makers operating in different languages. The situation doesn't significantly get worse when you operate an interlingua between language and language.

The initial situation, however, within Anglo-Nude, was already pretty bad. In order, however, fully to understand this,

* Lattite, Mark II, will however be supplied on request,

an account must be given of how to construct dictionary entries in Nude. And this account is relevant to the present discussion in any case, since Nude is the simplest interlingua any of us will ever see* (See overleaf)

It's obvious, from the above description of Nude that the elements of Nude are not in English. That being so, we can the more confidently compile Nue-France.

ELEMENTS OF NUE-FRANCE

No.	A-N.	N-F.	<u>No.</u>	A-N.	<u>N.F.</u>
0	not	non	1	bang!	conc!
2	done	fin	3	will	exprès
4	much	mault	5	for	but
6	cause	cause	7	change	change
8	can	peut	9	want	vent
10	laugh	rit	11	sense	sens
12	in	dans	13	have	a
14	pray	prie	15	use	util
16	do	fait	17	point	voilà
18	ask?	hein?	19	same	même
20	up	sur	21	think	esprit
22	feel	coeur	23	be	être
24	more	plus	25	whole	tout
26	count	nombre	27	one	un
28	true	vrai	29	please	bon
30	self	soi	31	part	part
32	folk	gent	33	man	homme
34	plant	plante	35	beast	bête
36	thing	chose	37	line	suite
38	world	monde	39	pair	pair
40	life	vivant	41	sign	signe
42	heat	feu	43	stuff	concret
44	grain	forme	45	kind	dorte
46	how	comme	47	when	quand
48	where	où			

<u>NUDE</u> <u>ELEMENT</u>	<u>APPROXIMATE AREA OF</u> <u>MEANING</u>	<u>EXAMPLE</u>
1. BANG!	Sudden action	bang:think (idea)
2. DONE	Completed Action	(done:change):folk (banquet)
3. WILL	Deliberate Intention	for:(will:do) (try)
4. MUCH	A Lot Of	have/(much;(count:(part:where))) (long)
5. FOR	Motive, Because	for:(will:do) (try)
6. CAUSE	Causative Actions	cause/(have/sign) (say)
7. CHANGE	Become	change/be (become)
8. CAN	Possible	(not can):(not have) (need)
9. WANT	Desire	(want/(cause/(change:not please))):man (enemy)
10. LAUGH	Humourous	laugh:sign (laughter)
11. SENSE	Senses and Perception	sense.SEE:heat LIGHT (light)
12. IN	Be Situated In, or having the Property of Being Able to Contain Something	in:thing (container)
13. HAVE	pertain "of"	cause/(nothave/life) (kill)
14. PRAY	Religious ideas	pray:man (priest)
15. USE	Appliances, tools etc.	((man/use):thing) (implement)
16. DO	Non-causative action	beast/do (animal actions)
17. POINT	Position in space	point:where (here, there)
18. ASK	Question, query	cause/have/(ask:sign) (question)
19. SAME	Similar, like, identical	not same (unlike)
20. UP	Elevation in space and society	(up:(part:folk) (aristocracy))
21. THINK	Cognition, know	true:think (know)
22. FEEL	Emotions	feel:not please (angry)
23. BE	Exist	change/be (become)
24. MORE	Increase, comparison	not more:line (end)
25. WHOLE	Complete	(count:man):whole (human race)
26. COUNT	Plurality, numbers	count:man(men)
27. ONE	Singular, same	one (same)
28. TRUE	Correct	true:think (know)
29. PLEASE	Satisfaction	feel:please (pleased)
30. SELF	Pertaining to oneself	(cause/(in/self)):stuff (food)
31. PART	Piece, or section.	part:folk (section of humanity)
32. FOLK	(Socially motivated) races	do;folk (custom)
33. MAN	Human Kingdom.	(part:folk):man (member of family)
34. PLANT	Vegetable kingdom	plant TREE (tree)
35. BEAST	Animal kingdom	beast/do (animal action)
36. THING	Inanimate object	(man/in):(thing: where) (house)
37. LINE	Sequence	not more:line (end)
38. WORLD	Pertaining to the physical world	up:(part:world) (sky)
39. PAIR	Pair	cause/(self :pair)/have,) (trade)
40. LIFE	Alive	cause/(not have/life) (destroy)
41. SIGN	Symbol(any sort)	cause/(have/sign) (speak)
42. FORCE	Energy	cause/(have/heat HEAT) (warm)
43. STUFF	matter	sign: stuff (money)
44. FORM	pattern(artistic,thought)	think:(stuff:grain) (chemistry)
45. KIND	Specie	same:kind (being of the same specie)
46. HOW	Mode, quality, adjective	think/same):how.
47. WHEN	Time	count: (part:when) (unit of time)
48. WHERE	Space	change/where (move)
49. SPREAD	Region in space or time	
00. NOT	Causes all Nude elements to mean their opposites.	

The syntactic devices of Nude consist simply of two connectives and a bracketting convention. The first connective, represented by a colon, stands between elements the relationship of which is that of adjunct and principle. The second, represented by an oblique stroke (/) is a non-commutative verbal connective representing the relationship of subject to verb or verb to object. To return to the example; it is clear that the connectives involved in "he says" are all verbal, and we write

man/cause/have/sign

In the case of "speaker", the idea of "speaking" is related to that of "man" as adjunct to principle, so that we write

man:cause/have/sign

Nude formulae are bracketted in such a way that every bracket contains two elements; these may either be primitive elements or other bracket groups. An adjunct is bracketted to the corresponding principle, an object to a preceding verb and a subject with the resulting predicate. The primitive elements are not themselves distinguished in respect of form-class; the syntactic relationships are expressed solely by the connectives and the brackets⁽¹⁾. The two formulae of our example thus finally become man/(cause/(have/sign)) and man:(cause/(have/sign.)).

The word-order is fixed for coding purposes; but the formulae tend to be written in the word-order natural to any dictionary-maker and then transformed into the standard order.

Any residuum of information, which the formulae will not accommodate, can be consigned to a list, one member of which is associated with each formula; provision is made for two list-numbers on each formula.

1. In Nude, as originally expounded by R.H. Richens (see "Interlingual Mechanical Translation", The Computer Journal, vol. 1 (1958) the connectives and brackets were represented orthographically by superscripts. This notation formally corresponds to the one described here.

The following kinds of divergence between Nude dictionary-entries have already appeared.

1. Plain chaos (i.e. nothing at all in common between different entries).

e.g. sacrament:

- i) thing: SACRAMENT
- ii) not thing: SACRAMENT
- iii) pray:sign: SACRAMENT
- iv) pray:grain: SACRAMENT
- v) man/feel/(pray:sign): SACRAMENT
- vi) man/sense/(pray:sign): SACRAMENT

In this case, an agreed preferred dictionary-entry can probably be obtained.

2. The two dictionary-makers are thinking of two different senses of the same word.

e.g. bank:

- i) Bank a) df. A place where money exchanges hands.
fmla. ((folk/do)/(have:(sign:stuff))): where BANK
- b) df. A geological shape (i.e. the shape of part of the world) (cf. SHAPE= df. (grain:where))
fmla. ((part:world):(grain:where)) BANK
- ii). Change a) df. a particular change, or kind of change.
fmla. (change:kind) CHANGE
- b) df. to change something into something else, or for something else.
fmla. cause/change CHANGE
- c) df. "To change", used intransitively.
fmla. change CHANGE

In this case, a separate card can be key-punched for each entry.

3. The two dictionary-makers produce only slightly differing Nude entries, probably through one of them carrying the Nude analysis much farther than the other does.

In this case, the two entries should be conflated. If the resultant entry is too long, it will usually be because two senses of the word have been described simultaneously; the entry should then be split, and two cards made.

4. Monolingual (i.e. individual) prototypic formulae appear in Nude.

This is the non-corrigible case of dictionary-makers divergence since it is equivalent to Nude-developing idioms. Inevitably, in a system with so few elements, combined elements will tend to have individual meanings e.g. (sign:stuff): MONEY. The trouble arises when dictionary-makers make these by using Nude elements in new ways. Consider, for instance, the Nude element UP! The original function of this was to indicate elevation in space. But now take (up:(part:folk)):man:NOBLE. Here UP! clearly means elevation in society; a metaphorical use of UP! has been made.

This is disturbing in two different ways. Firstly, an intersection between a word with a literal UP! and a word with a metaphorical UP! will not be a genuine intersection in Nude. Secondly, the only way of warning the machine of this is by preparing the metaphorical UP! by another, and special, Nude prototypic formula (e.g. (not same: up):(up:(part:folk)):man.

Experience shows, however, that this device may not save you. For, - especially when C.L.R.U.members start talking in Nude, making jokes in Nude and writing letters in Nude, - the new prototypic formula may itself become an idiom (i.e.(not same) may itself begin to be used in varying ways); and so on. Thus Nude will develop more and more of the idiomatic (i.e. non-interlingual) picturesqueness of a natural pidgin, as opposed to the invariance required of an artificial pidgin language.

So it comes to this of the four ways in which dictionary-making divergencies emerge in Nude; the first three can be rendered interlingual, but, unless we can invent some device to deal with it, the fourth cannot. By using mechanical aids, we can track what happens when Nude idioms develop; but we cannot stop dictionary-makers from developing them. And this brings up again, undeferably this time, the up to now submerged question: how much does it matter if idiomatic meaning in Nude is rendered arbitrarily?

* In reaction from this, Lattite is being liberally sprinkled with metaphor-markers, in the hope that, if given enough, people will not use these metaphorically.

It becomes evident here, too, just why it does matter; why, in fact, it has mattered all along. The purpose of an interlingua is to operate interlingually; that is, to be a vehicle of translation between many languages. Now, when bi-lingual interlingua-using translation is contemplated, exact matches can be arranged between arbitrary formulae. When interlingual translation is contemplated, they can't: you cannot sufficiently keep track of the formulae in all languages. Moreover, the over-frequent occurrence of arbitrary prototypic formulae takes all effectiveness from the real, catastrophic time-saving procedure which an interlingua permits. This is, not that you should make a mere n interlingual dictionaries instead of $n^{(n-1)}$. It is that having made and tested one good interlingua, you can, from it, mechanically generate all the others. For an interlingual punched-card pack can be mechanically reproduced; and mechanically interpreted into the Nude of the new language.* But it is still necessary for the second language's dictionary-maker 1) to assign stretches of his native language to each of the Nude formulae in his reinterpreted pack, 2) to add, and make a Nude formula for, and key punch extra cards for, any frequently used word in his language not allowed for by the pack.

Now, it is a matter of experience that it is extremely easy, with practice, to make a Nude entry for a word; but extremely difficult to find a word for a Nude entry. (In the first enterprise, you have only 48 elements, and two connectives to choose between at each stage of the compilation; in the second enterprise, you have the subwords, words and phrases of the whole language.) Two helps are available. One is to look up the translation of the word you want in the Anglo-Nude Glossary; a special glossary which is being compiled for this purpose. The second is to use a thesaurus in the second language; and then to identify the rows and lists, not the rows, by Nude formulae. Both of these devices are in practise unusable, however, if arbitrary Nude formulae have been used in the first place; for there will be not merely arbitrary, but gibberish, in the second language.

And this fact, that arbitrariness is inadmissible in Nude, should make us see that it always has been, all along. Arbitrary heads, arbitrary-syntax markers, archeheads, row specifications, none of them will do once you handle more

* That is, the engineering difficulties of interpreting Marcode can be overcome.

than two languages. And this should lead us to recapitulate the argument of this section: to ask ourselves: "How much, in fact, were the heads and the rows arbitrary, and how much interlingual?"

As from now, my conclusions on this point are threefold.

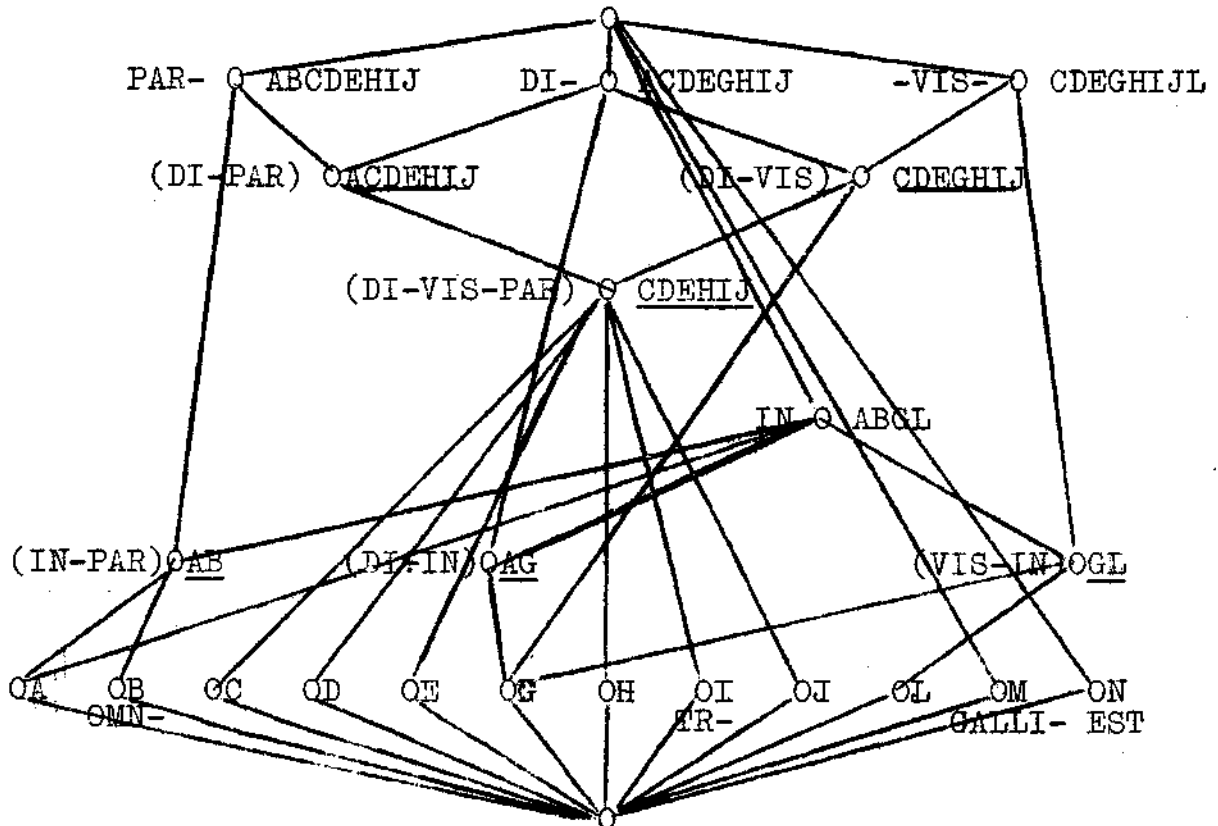
1. If the arbitrary heads in Roget (as shown by King's test) are dropped, and an overlapping set of heads are chosen from the thesauruses which did not have the same set of heads in the first place, a set of interlingual heads can be found. (This is partly because corresponding heads can have different, lattice "areas", since the unit of a head, in a natural thesaurus, is a vague one.)
2. Rows, which form a much sharper unit, may not be interlingual, but the components by means of which they are specified must be.
3. It is vital that, whatever set of row-specifiers are chosen, they themselves should not be used metaphorically.

END OF IV

V. THE CONTEXTS OF A SENTENCE SEEN AS A SUB-LATTICE
OF A THESAURUS

In the Introduction to this paper, I gave a definite undertaking to write this Section unsatisfactorily. This undertaking, I will now proceed to fulfil.

Consider the lattice given below:



This was obtained by the following procedure:*

1. The heads occurring most frequently in the stem of the words of "Gallia est omnis divisa in partis tres", that is to say, the heads occurring in the dictionary entries of 'Galli-', 'est', 'omn-', 'di-', '-vis-', 'in', 'par-' and 'tr-'.
2. All the heads occurring more than four times were retained.
3. The retained heads were then structured into a lattice with the individual heads as minimal elements by the following procedure. The top point of the lattice is a latent element which is the join of all the minimal elements in the lattice. The chunk, or chunks, which have the greatest number of heads from the highest rank of the lattice, and the other chunks below are related to this rank, to the top, and to each other by the inclusion-relation which defines the nature of the lattice, and which is as follows: any element ABC includes the elements AB, AC, and BC, and also A, B, and C, these last three being minimal elements. There

* This procedure was carried out by C. Wordley. The dictionary entries had been made by MM and KSJ for quite other purposes.

may be latent elements representing heads which are common to chunks, but these head sets do not represent actual chunks. The bottom point of the lattice is the meet of all the elements.

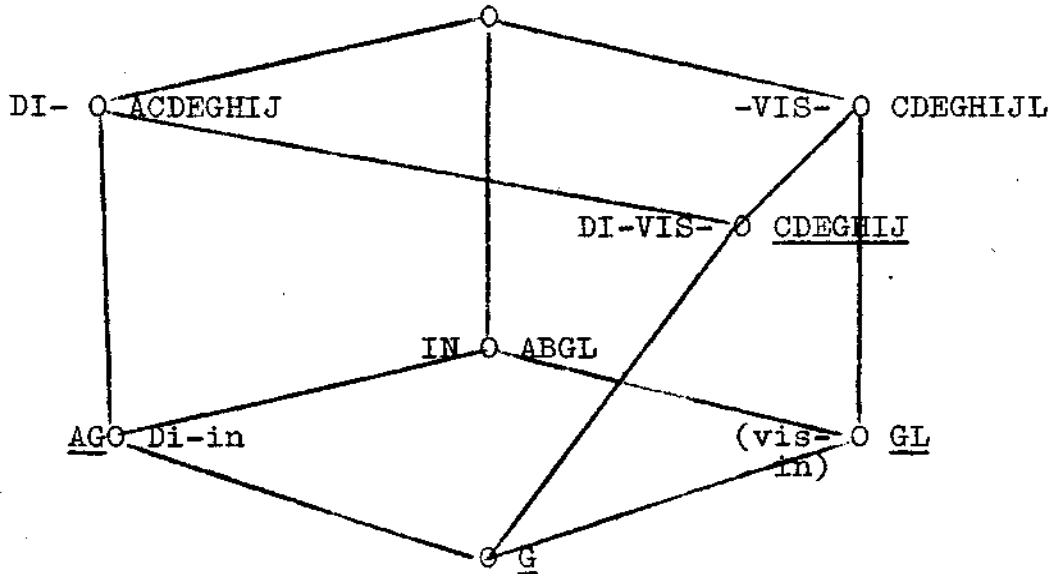
The resultant lattice will always be derivable from a sub-lattice of the total Thesaurus.

This lattice was highly interesting to us, for the following reason. In 1956-57, Masterman and Parker-Rhodes, in the course of various unsuccessful attempts to construct a non-contentious interlingual syntax-lattice which should operate independently of the semantic part of the thesaurus, had formulated the following four criteria of success in applying lattice-theory to this problem.

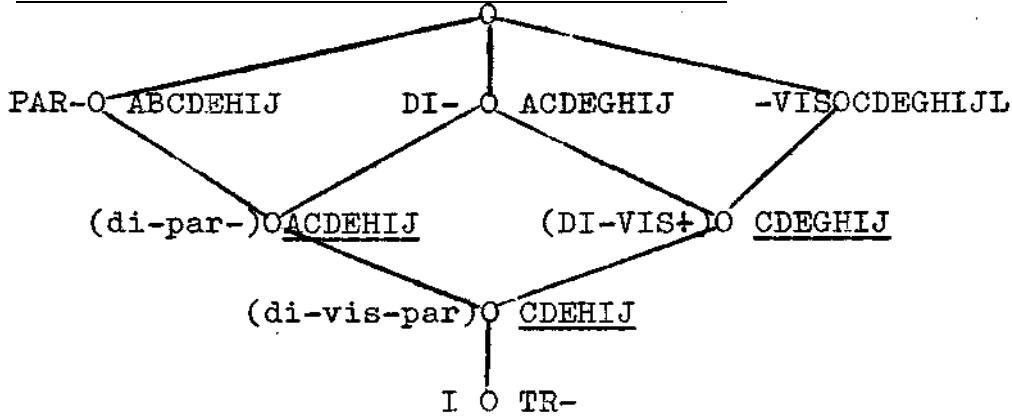
1. There must be a direct lattice connection between every adjective and every noun, and the adjective including the noun.
2. The subject and the predicate must form separate sub-lattices. If the subject were single, it must therefore connect straight with the I element.
3. Prepositional phrases must form sub-lattices.
4. Words in the text joined by 'and' must form a Boolean diamond with the 'and' as the join.

Now, the lattices given below, which were constructed with no thought in mind of conforming to these criteria, in fact conform to the three which are relevant very well indeed. (Overleaf)

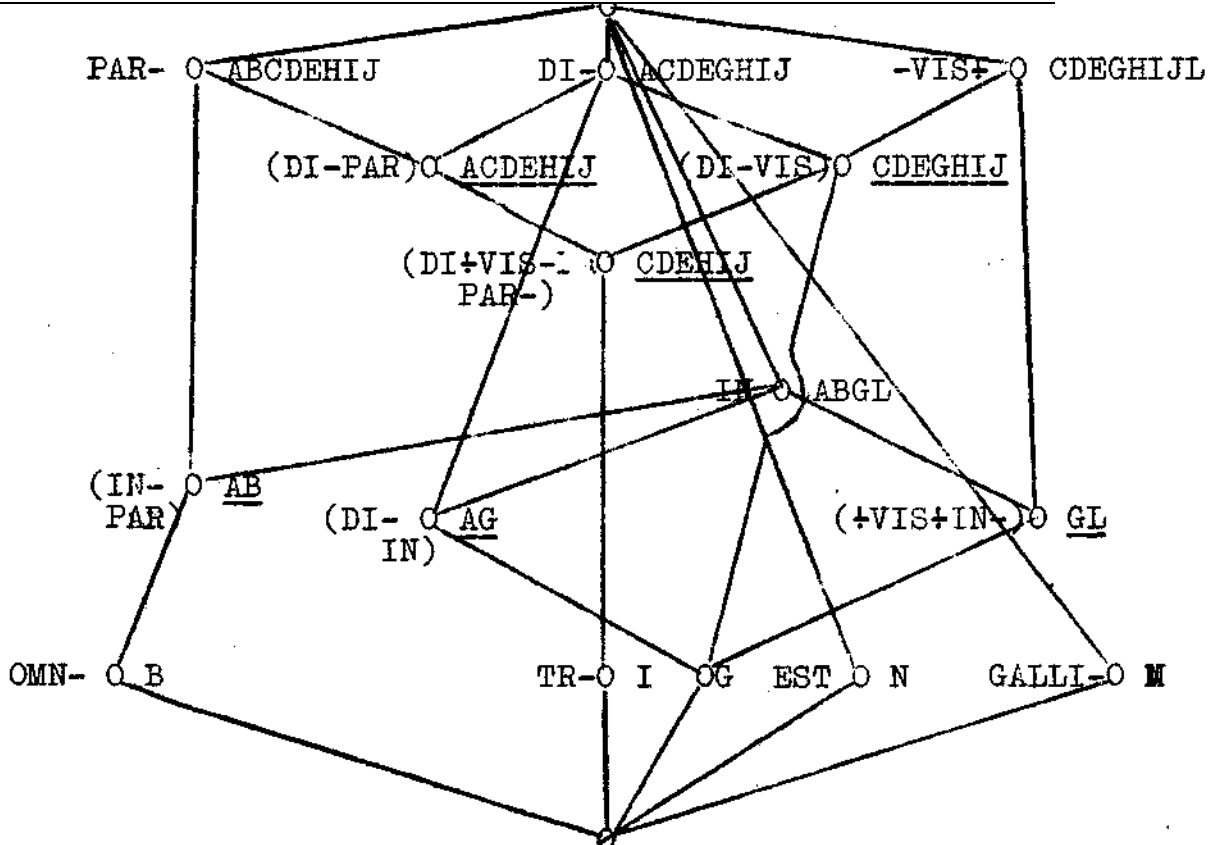
Sub-Lattices taken from the "Gallia" Lattice
Sub-Lattice 1. Divisa in.



Sub-Lattice 2. Divisa in partis tres.

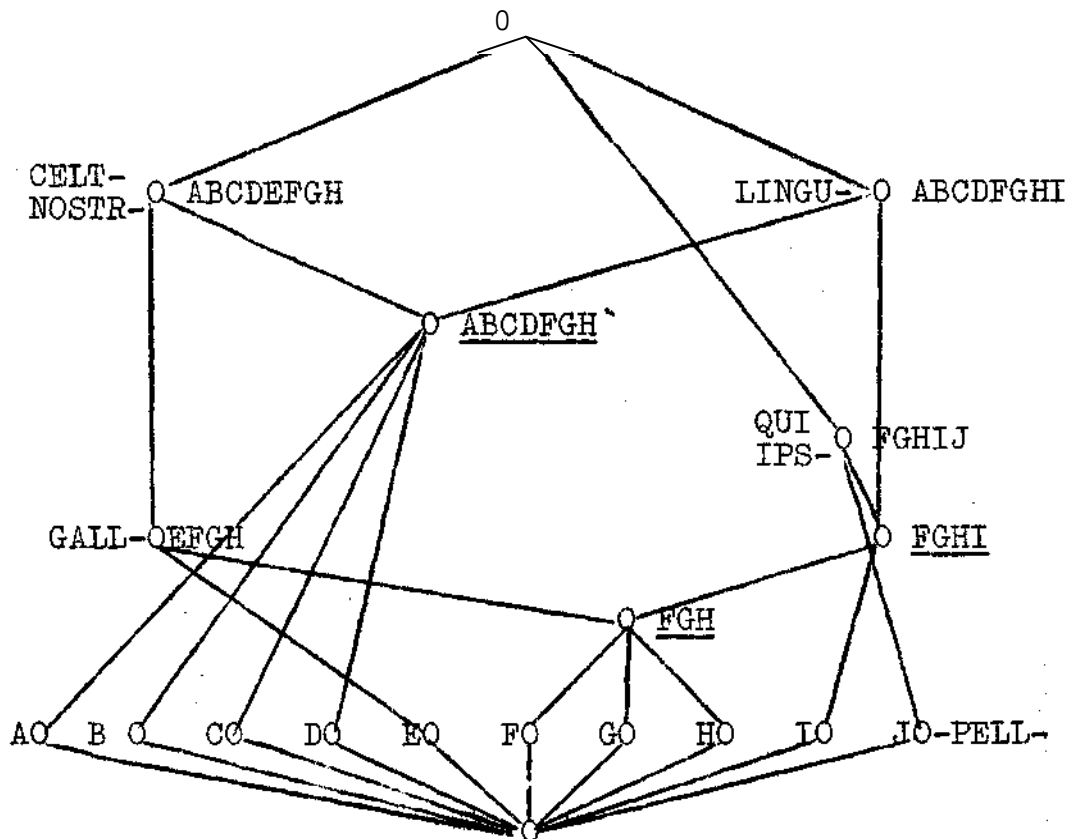


Sub-Lattice 3. Showing subject and predicate division.



This suggested the possibility that context structure thus defined has sentential significance. Further trials*, however, falsified this claim. Consider, for instance, the sentential shambles produced by the same procedure in the lattice below:

The sentence considered is, "Quarum unam inclount Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur."



Clearly the procedure in its present form is not right. Nevertheless, it is very difficult, having once used this procedure, to refrain from continually tinkering with it to try and make it work better. For that there is a well-marked class of frequently occurring heads in any sentence is established, and that a lattice can be made from these heads by the procedure, is true, too. The question that arises is, therefore, how to devise an analytic procedure which makes use of these facts. The following developments immediately suggest themselves;

1. To combine this programme with a bracketting programme, so that the lattices were made with clauses, not sentences.

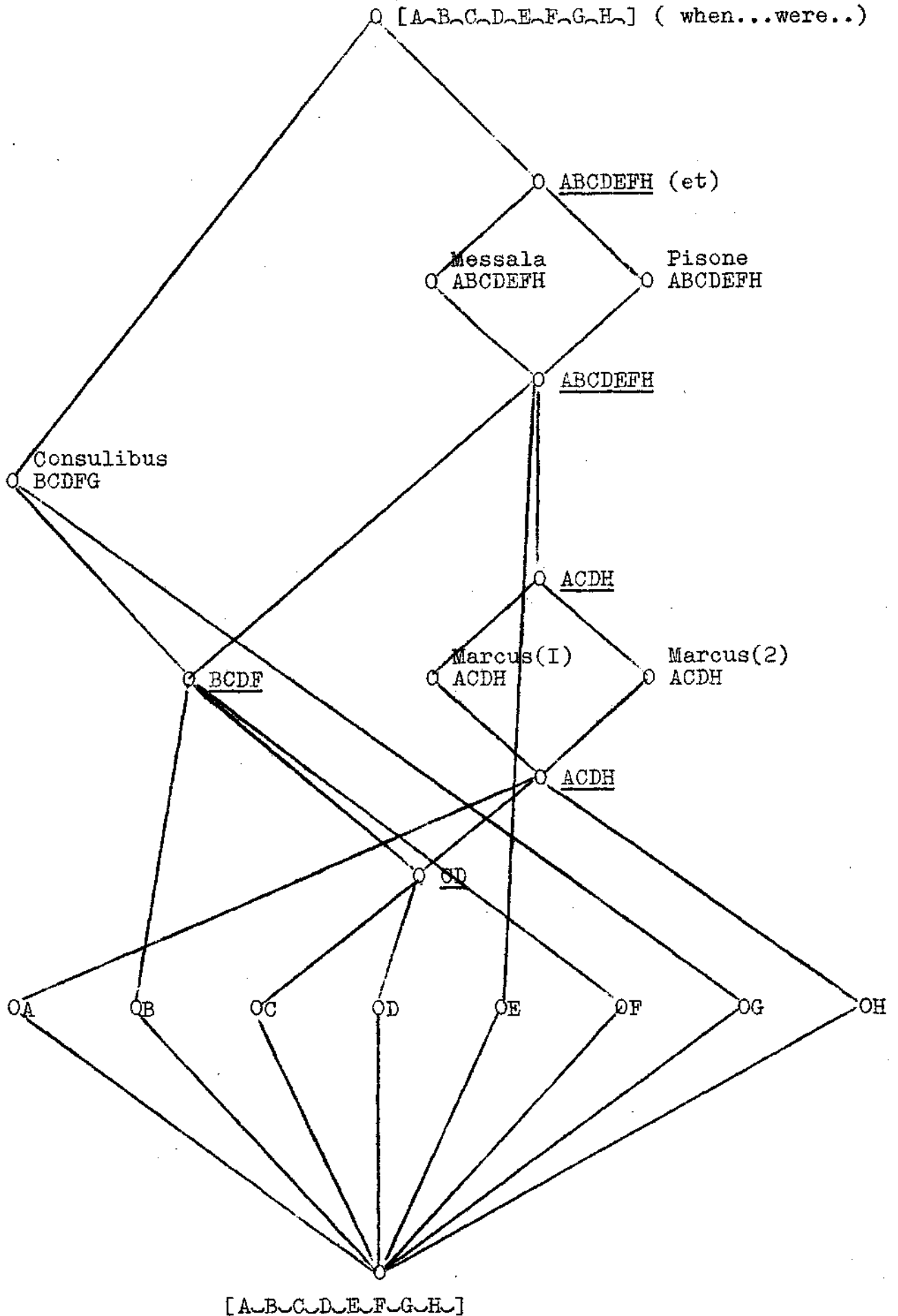
* These were made by C. Wordley over 15 sentences and parts of sentences.

2. To regard the set of syntax-markers of the sentences as an extra set of frequently occurring heads. For instance, the syntax-markers could be envisaged as heads occurring frequently not in the sentence but in the language.

3. To construct, intuitively, "model" context-structure lattices, using developments 1 and 2, and conforming to the criteria; and then try to reproduce these mechanically.

An example of such a "model" lattice is given overleaf:

".....Marcus Messala et Marcus Pisone consulibus....."



A Singular
B Plural

E Feminine
F Time When

C Human
D Masculine

G Rank
H Proper Name

This line of research, as the above account makes clear, has so far only been scratched. Its value, however, even at this stage, seems to me to be that it indicates the possible existence of two types of sentential patterning. The first of these, which might, in the end, turn out to be an emphasis pattern, is formed by using lattice-theory to relate the most frequently occurring contexts. The second, which it is not known how to handle yet, is formed by performing bracketting and other operations upon the syntax-markers given by the dictionary entries.

The question is, can each be controlled, so as to become empirically significant, and then the two related? Or is each an arbitrary structure thrown up by performing operations which are congenial to the theory, and therefore of no significance in actual language?