6. The potentialities of a mechanical thesaurus

There are six sections to this chapter:

1. The logical effect which adopting the logical unit of the MT chunk, instead of the free word, has on the problem of compiling a dictionary.

2. Dictionary trees: an example of the tree of uses of the Italian chunk PIANT-.

3. Outline of a Mechanical Translation-programme using a Thesaurus.

4. Examples of trials made with a model-procedure for testing this: translations of ESSENZ-E, GERWOGL-I and SI PRESENT-A from the Cambridge Languages Unit's current pilot-project. The simplifications which the use of a Thesaurus makes in the research needed to achieve idiomatic Machine Translation.

5. Some preliminary remarks on the problem of coding a Thesaurus.

1. In MT literature: it is usually assumed that compiling an MT dictionary is, for the linguist, a matter of routine; that the main problem lies in providing sufficient computer-storage to accommodate it. Such judgements fail to take account either of the unpredictability of language, (Reifler, 1956), or of the profound change in the

conception of a dictionary produced by the substitution of the MT *chunk* for the *free word*.

By *chunk* is meant here the smallest significant language-unit which(i) can exist in more than one context, and (ii) which, for practical purposes, it pays to insert as an entry by itself in an MT dictionary. Extensive linguistic data are often required to decide when it is, and when it is not, worth while to enter a language- unit by itself as a separate chunk. For instance, it has been found convenient to break up the Italian free word *piantatore* into the chunks PIANT-AT-ORE. It has not been found convenient to break up the Italian free word *agronomi* into chunks AGRO-NOM- I, but only into the chunks AGRONOM-I, since the addition of -NOM-to-AGRO- enables the distinction to be made between AGRO-, meaning "agriculture", and AGRO-, meaning "bitter".

Experience shows that the cutting-down of the number of entries, and the compensatory extension of the range of uses of each entry, caused by the substitution of chunks for free words, are together sufficient to call in question the current conception of a dictionary-article. In this paper we shall speak of current dictionary-*articles*, MT dictionary-*entries*, and Thesaurus *items*.

2. From the logical point of view, it can be shown that the range of uses of any chunk form a tree. Some paths of this tree are open to alternative analysis, but a considerable number of the paths, as of the points, can be determined on objective criteria determined by the immediate context. For instance, the use of the Italian chunk PIANT- in the free word *piantatoio* is clearly different from its use in the free

word *piantatore*. Moreover, the design of the tree can often be tested by its predictive value; for instance, in making the tree of the chunk FIBR-, a junction-point which had to be inserted to account for well-established uses was later found, when a larger dictionary was consulted, to be exactly fitted by the use of FIBR- in the free word *fibroso*, which had not appeared as an article in the smaller dictionary. [….]

Dictionary-articles which contain PIANT-

1. im-PIANT-ament-o, s.m., impiantation, building, establishment.
2. im-PIANT-are, v.tr., to establish, to settle down to business, to found. impiantare una scrittura, to open an account.
3. im-PIANT-arsi v.rifl., to take one's stand.
4. im-PIANT-it-o s.m., floor, tiled place.
5. im-PIANT-o, s.m., establishing, setting up of a business.
6. PIANT-a, s.f., plant; tree; (arch:) plan, groundwork; sole (*pianta dei piedi*); lineage i.e. family tree: fig race: *pianta esotica*, exotic plant; *pianta di un edificio*, plan of a building; *essere in pianta*, to be on the list, *rifare una cosa di sana pianta*, to do a thing a second time.
7. PIANT-abil-e, adj., pertaining to a plantation.
8. PIANT-aggin-e, s.f., plantain, i.e. pasture-plant.
9. PIANT-agion-e, s.f., plantation, planting: *piantagione di patate*, potatofield.
10. PIANT-ament-o, s.m., planting, plantation.

11. PIANT-are, v.tr., to plant; to set; to stick; to drive in; to place; to forsake, to abandon, of, French plaguer. *piantare una bandiera*, to set up a standard, to hoist a flag. *piantare in asso*, to leave a person in the lurch. *piantare un pugnale nel petto*, to stab with a dagger of English, stick a dagger into him. *piantare carote* (fig) to make someone believe of English, to plant a clue *piantar le tende*, to lodge, to dwell.

12. PIANT-arsi v.rifl., to fix oneself, to settle: to set up; *piantarsi in un lucco*, to settle: to set up; *piantarsi in un lucco*, to settle down in one place.

13. PIANT-a-stecch-i, s.m., (calz.) punch, puncheon (*arnese per piantar gli stecchi nelle suole*).

14. PIANT-at-a, s.f., plantation: row of trees.

15. PIANT-at-o, part.pass.e.adj.planted, set up: fixed: *ben piantato*, well-built, well-set-up man.

16. PIANT-at-oi-o,m., (agr.) tool for planting, dibbler.

17. PIANT-at-ore,-trice, s.m.f., planter.

18. PIANT-at-ur-a, s.f., plantation, planting.

19. PIANT-im-i, s.m., plur., many sorts of plantations.

(PIANTO, s.m., tears, weeping; lament; (fig) pain; regret.)

(PIANTO, part, pass., wept; lamented; deplored.)

20. PIANT-on-ai-o, PIANT-on-ai-a, s.m.f., (agr.) nursery.

21. PIANT-on-are, v.tr., to watch over, to nurse, to guard: to piant outtings.

22. PIANT-on-e, s.m., (mil) sentry nurse, guard. (fig) watcher; (agr.) sucker scion, sapling. *Essere di piantone*, to sentine, to be on !! guard, to guard.
23. s-PIANT-a-ment-o, s.m., uprooting, transplanting.
24. s-PIANT-are, v.tr., to uproot, to transplant; to ruin.
25. s-PIANT-at-o, s.m., penniless person; (fam) someone who is dead broke, opp. stony broke.
26. s-PIANT-o, s.m., ruin, destruction. *Mandare a spianto*, to ruin.
27. tra-PIANT-a-ment-o, s.m., transplantation.
28. tra-PIANT-are, v.tr., to transplant.
29. tra-PIANT-at-oi-o, transplanter (a tool).

When the contexts provided by translation into a second language are added to the above, the tree becomes very much more complicated. Inspection immediately shows, moreover, that the only criterion for differentiating many of the new points on the bi-lingual tree is the fact that, if, say, of two otherwise similar uses of PIANT-, English translations are given, different English words will be used in the two cases. For instance, once the English language is considered as well as the Italian, the use of PIANT-in the phrase *piantar le tende*, "to pitch a tent", must clearly be distinguished from its use in the phrase *piantare una bandiera* "to set up a standard". But to the man thinking wholly in Italian, this difference of use may not be perceptible: for him, one plants a tent on the ground and a standard in the air in exactly the same figurative sense of "plant"; all the more so, indeed, as *piantar le tende* means permanently to establish a tent (compare "Caesar then established his winter quarters") and is to be contrasted with *rizzare le tende*, which means to pitch a tent with the intention of

taking it up again in a short time, - and this last differentiation of context is one which we have not got in English.

Such considerations raised doubts of the validity of such *translation-points* on bi-lingual dictionary-trees, which led to the re-analysis of bi-lingual dictionary-tree not as trees but as lattices. For translation-points on a dictionary-tree are not just points on a single path but junctions of two paths; as, indeed, the contexts of the uni-lingual tree might also be taken to be if such chunks as -UR- and -AGION- were taken as the points of origin of trees. Moreover, if it be granted that, even in simultaneous translation, translation is never actually made between more than two languages at once, a multi-lingual tree, as opposed to a bi- lingual tree, will also have this property that all its points will be translation-points, and it will therefore be a lattice. Moreover, it will not always be true that as the number of languages which are incorporated increases this lattice will become significantly more complex, because many of these translation-points will fall on one another. [….]

3. In this design the chunks of the input-text are passed through four successive processes of transformation. The first stage of each of these consists of matching the chunks, in turn, with some sort of dictionary; there are thus four dictionaries used in succession in the programme. These are 1. the bi-lingual pidgin- dictionary: 2. the lattice inventory: 3. the thesaurus cross-reference dictionary: 4. the thesaurus.

In order to exemplify this whole mechanical-translation process in concrete form, the following test-procedure has been devised. Translation trials might be undertaken

which, if MT is to develop as a subject in its own right, will provide the controlled empirical material which we so much need.

In the procedure described below, the lattice-inventory and programme, which is by A.F. Parker-Rhodes, will in the near future actually go through a computer, The Thesaurus used was Roget's Thesaurus, (1953 edition), amended and amplified according to the procedures given below. The general design was by Masterman, and the pidgin passage-dictionary by Masterman and Halliday. The matches were made by means of alphabetically-stacked packs of written cards, each containing the entry for one chunk. In the case of each pair, was stacked first.

Since the method of matching with the lattice-inventory is more complicated, an appendix explaining the chunk-interpretation of lattice-theory as it is being used in the Cambridge Language Research Unit, and made with special reference to the Italian paragraph which is used to illustrate the Thesaurus test-procedure, is attached to this paper.

The procedure was developed as follows: A paragraph from an Italian botanical paper was chosen, and divided into chunks as given below:

LA PRODUZ-ION-E DI VARIET-A DI PIANT-E PRIV-E DI GEMM-E ASCELL-ARI , O PER+LE+MENO CON GERMOGL-I A SVILUPP-O RIDOTT-O , INTERESS-A DA+TEMPO GENET-IST-I ED AGRONOM-I, . TAL-E PROBLEM-A SI+PRESENT-A PARTICOLAR-MENT-E

INTERESS-ANT-E PER ALCUN-E ESSENZ-E FOREST-AL-I E FRUTT-IFER-I, PER LE PIANT-E DI FIBR-A, MA SOPRATTUTTO PER IL TABACCO-O. IN QUEST-A COLT-UR-A E INFATTI IMPOSSIBIL-E MECCANIZZ-ARE L'-ASPORT-AXION-E DEI GERMOGL-I ASCELL-ARI , NECESS-ARI-O D'+ALTRA+PARTE PER OTTEN-ERE FOGLI-E DI MIGLIO-E QUALIT-A. .

(N.B. Entries of the form A+B+C..+N were entered as single chunks)

A simple Italian-English pidgin dictionary was then compiled covering the chunks of this paragraph. Specimen entries taken from this are given below. It was be noted that while the schema of this dictionary allows of one chunk having, if necessary, two Lattice Position Indicators, (L.P.Is), though the chunks entered in this passage-dictionary have only one, it does not allow of any chunk having more than one pidgin translation. The numbers in the right-hand column govern a very simple first-approximative procedure for assigning singulars and plurals. The whole passage-dictionary was planned to give, as simply as possible, an output embodying only what the machine could immediately find out of the structure of Italian.

Sample Italian-English Pidgin-Dictionary Entries

Italian	L.P.I.	1→0 routine	English
-A	28		ω
-AL-	39		-Y

DA+TEMPO	28		FOR+SOME+TIME+PAST
FIBR-	30		FIBRE
I	26	0→1, 1→1	THOSE-WHICH-ARE
GENET-	60		GENETIC-

When the chunks of this dictionary were matched with the chunks of the input, the following output was obtained:

Output I: top line: singular/plural subroutine

second line: output in chunks

decimal numbers: L.P.I.s

initial set of subroutine (i.e. unmarked form) 1

INSERT FIGURE 18 HERE

The L.P.I.s of the chunks of this output were then picked up and inserted uniquely into lattices by means of the lattice-inventory and lattice-programme (see appendix). These lattices give synthesis-routines for english which produce output II, below:

Output II:

THE PRODUCE-MENT OF VARIETY-S OF PLANT-S WITHOUT AXIL- ARY BUD-S, OR AT+LEAST WITH SPROUT-S AT REDUCED DEVELOPMENT-S, (SING) INTEREST-(PRES) FOR+SOME+TIME+PAST GENETIC-IST-S AND AGRICULTURE-IST-S (PLUR). SUCH PROBLEM- S (PLUR) SELF-PRESENT

(PRES) PARTICULAR-LY INTEREST-ING FOR SOME FOREST-Y AND FRUIT-BEARING ESSENCE-S, FOR THE PLANT-S OF FIBRE-S, BUT ABOVE ALL FOR TOBACCO. IN THIS CULTIVATE-URE IT BE (PRES) IN+FACT IMPOSSIBLE TO MECHANIZE REMOVE-MENT OF+THE AXIL-ARY SPROUT, ON+THE+OTHER+HAND NECESSARY FOR TO OBTAIN LEAF-S OF BETTER QUALITY-S (PLUR).

It will be noticed that, in this output, the translation-procedure fails for non-grammatical reasons at a few easily identifiable points. (I am ignoring spelling-mistakes produced by the pidgin, such as PRODUCE-MENT for "production", as these could be picked up by cross-entries in the Thesaurus cross-reference dictionary). ESSENZ-E, in the original, is translated ESSENCE-S; GERMOLGL-I is translated SPROUTS; SI PRESENTA is translated SELF-PRESENT; and if ASCELL- has been given its vernacular meaning of ARMPIT-, the phrases ARMPIT-ARY BUD-S and ARMPIT-ARY SPROUT-S would have occurred in the translation.

In order to decide between AXIL- and ARMPIT-, as the translation for the pidgin-dictionary, a trial was made by rendering into pidgin the biblical story of Jeremiah the prophet, who was rescued from the pit by ropes which rested on the rags which he had put under his axils. This story remained comparatively comprehensible. This result could semantically, have been foreseen, since an armpit is an instance of an axil, as is also the crutch of the legs - the only other place Jeremiah could have put his rags, whereas the idea of an axil cannot, inductively, be reached from that of an armpit.

It was therefore decided further to examine these cases, by putting them through the Thesaurus cross-reference-dictionary and the Thesaurus.

Roget's Thesaurus cross-reference dictionary is arranged alphabetically. The entries in it form trees, but much simpler trees than those produced by normal dictionary-entries. Specimen entries from it are given below:

Specimen Entries from the cross-reference-dictionary of Roget's Thesaurus

bud 367	plant	problem 454, 461, 533,
beginning 66* 129*	place 184	-atical 475
erm 153	insert 300	
ornament 847	vegetable 367	
expand 194	agriculture 371	
graft 300	trick 545	
-from 154	tools 633	
-dy 711, 890	property 780	
	-a battery 716	
	-oneself 184	
	-ation 184, 371, 780	

It will be noticed that into the specimen entries given above, cross-references (between asterisks) have been inserted in the entries for *bud* and *problem* but not in the entry for *plant*. These insertions have been made to make the Thesaurus multi-lingual. They have not, however, been made *ad hoc*. If the Thesaurus dictionary

procedure given here is to work for translation-trials, additions and emendations to the Thesaurus must be made only according to Thesaurus-principles; that is, according to one of the procedures given below:

Procedure for amplifying a translation-Thesaurus

Each chunk in any pidgin-dictionary must successfully match with an entry in the cross-reference-dictionary: e.g. PLANT-, *plant*.

Each main meaning of the corresponding source-language entry in the pidgin dictionary must be compared (*not* matched) with the sub-headings of the cross-reference entry. If the comparison is unsatisfactory in that there is reason to suspect that the cross- reference-spread is too narrow, (i.e. that the cross-reference-tree has not enough main branches) then one of the two emendation- procedures given below must be adopted.

(i) without making an addition to the cross-reference entry, bring down the actual Thesaurus-items which are referred to in the entry and search for the missing meanings. If they are found, no addition to the cross-reference entry need be made.

Example: The Italian bi-lingual dictionary tree of PIANT- (actually a lattice) has a branch with the main meaning *design*. This branch has derived meanings *groundwork, plan, blue-print, installation; list; scheme, invention, pretext, lie*. In the cross-reference entry ‘plant as design' does not occur. “plant as trick, 545”, however, does; and the Thesaurus item 545, *Deception*, gives, either directly or by sub-

reference, *lie, pretext, invention, and blue-print*. *Scheme, design* and *plan* can also readily be reached from this item if (under emendation-procedure (ii) below an addition is made to item 545, row 3, so that this row now reads:

item 545, row 3: *trick, cheat, wile, ruse, blind, feint, plant, catch, chicane, juggle, reach, hocus; thimble-rig, card-sharping, artful dodge, machination, swindle, hoax, hanky-panky; tricks upon travellers; confidence trick; strategem, &c. 702; scheme, &c, 626, theft, etc., 791*.

That the new asterisked element is a legitimate addition to the Thesaurus can be confirmed by consulting item 702, *Cunning*, where *schemer* occurs, and where there is a reference back to 545.

"*List*" could legitimately be inserted into 626 as follows:

item 626, row 4: List, programme, &c, 86; forcast, play-bill, prospectus, scenario, . . .

This addition can be checked by looking up 86, *List*, which already contains *programme*.

List should also be inserted into 626 in row 11, so that this now reads:

item 626, row 11: *cast, recast, systematise, organise; arrange, list &c, 60, digest, mature*.

This addition can be confirmed by consulting item 60, *Arrangement*, which already contains *list*.

Finally, under *list*, in the cross-reference dictionary, a subheading must now be added "list as plan, 626", so that the total entry now reads:

list

as catalogue 86

as plan 626

as strip 205

as leaning 217

etc. etc.

Of the remaining meaning of PIANT-, routes to *groundwork* and *installation* can only be constructed, it at all, by more intervening steps, since the items 25, *Support* and 185, *Location*, where they occur, do not appear in the dictionary cross-reference entries of any of the others.

Thus there is no incentive to add "plant as design" to the cross- reference-entry of *plant*, which would be done under procedure (ii) since the entry "plant as trick" already leads to all the items which could thus be reached.

(ii) Under this procedure an addition is made to the actual cross- reference dictionary-entry of the chunk in question.

Example: The bi-lingual dictionary-tree (actually a lattice) schematising the uses of GEMM- contains a branch of which the main English meaning is "*gem*". The

Thesaurus cross-reference dictionary-entry BUD includes no cross-reference which leads to any item containing *gem* or *jewel*. If, however, the cross-reference, "bud as ornament 847" is added between "bud as germ" and "bud as expand", (see above), the required connection is made, since item 847, *Ornament* reads as follows:

item 847, row 7: *tassel, knot, epaulet, frog: star, rosette, bow: feather, plume, aigrette*.

row 8: *jewel: jewellery; bijouterie, diadem, tiara: pendant, trinket, locket, necklace, armilla, bracelet, bangle, armlet, anklet, ear-ring, nose-ring, chain, chatelaine, brooch*.

row 9: *gem, precious stone; diamong, emerald, onyx, plasma; opal, sapphire, ruby; amythyst, pearl...*

We now have the required connection from entry to item. In order to be able to get back from item to entry, however, one of the given rows of 847 must be extended so as to include *bud*. The suggested extension is as follows:

item 847, row 7 (contd.)...*feather, plume, aigrette; bump, button, nipple, nodule, bud*.

The justification for this extension, of course, has got to be that some, at least, of this chain of metaphoric uses exist in English. *Bump* can be taken as colloquial: ("that is a very ornamental bump you have upon your forehead".). Ornamental *buttons* are dressmaking stock in trade; this element should be already in the item. *Nipple* has a definite, though rare, use as a nipple-shaped beautiful object. ("The crests or nipples

of the hill line are crowned with the domes of the mosques", wrote Cory in 1873: Oxford Dictionary). *Nodule* has an even rarer one, meaning "something like a knot". Finally, *bud*, meaning "ornament" does exist, but only poetically and archaically. Thus we get "Their breasts they embuske on high and their round Roseate buds immodestly lay forth", (Nashe, 1613). And Emerson, in his poems, wrote much later of "the bud-crowned spring".

Thus we get the curious situation that the use of an extended train of meanings for *ornament*, all of which have become cliches in Italian, is still an act of poetic originality in English and American. Nevertheless, the train of uses exists, and the addition to the Thesaurus item is therefore justified.

These methods of emending and amplifying Roget's Thesaurus have been exemplified in detail, because, in view of the surprisingly good outputs which follow, it might be thought that the Thesaurus-routes used had been manipulated to suit the Italian paragraph. This is not so; every suggested new connection has been checked and justified, and all relevant asterisked emendations used to reach the outputs are given below. The suspicion of manipulation represents a direction opposite to that in which the research has gone, for, in actual fact, the more the experience which is gained of using this Thesaurus, the less the emendations which are made. It is a sound presumption that, with few exceptions, all possible chains of meanings are somewhere in Roget's Thesaurus if they can be found. A minimum number of trials, moreover, begets a strong conviction that Thesaurus searching and matching would best be done automatically from the earliest possible date; they are no work for a mere human being. In other words, if the thesaurus-technique proves, on trials, to have definitive

MT significance, it will also prove to be the frontier-point where the MT worker, in this new kind of calculation, hands over to the machine; where results, uncalculated in advance by the programmer, are produced by the programme. It may also (that is, if it establishes itself as having translation value), be the point of departure for a new exploration of the analogy between the human cortex and a computer; for this feels like a model of what we do when we ourselves translate.

4. Work done on the Italian paragraph has provided the following examples of translations produced by the Thesaurus procedure.

Case I: ESSENCE-S.

If the chunks FOREST AND FRUIT-BEARING ESSENCE-S, - that is, all the chunks in the invertor-lattice 56, 60, 60, in which they occur (see appendix) are matched with the entries in the thesaurus cross-reference dictionary, the following output is obtained:

Output III:

forest 57 367, 890 and 37, 38 fruit result 164

produce 161

food 298

profit 775

forbidden- 615

reap the -s, 973

-tree, 367

fruitful 168

fruition 101

fruitless 169, 645, 732

bearing

relation 9

support 215

direction 278

meaning 516

demeanour 692

-rein 752

fruit-168, 637, 367

child-161

essence 5, 398

essential

intrinsic 5

meaning 516

great 31

required 630

important 642

essentially 3, 5

essential stuff 5

Upon this output the thesaurus-operations are performed with the aid of restrictive and permissive rules, given as they occur, and the object of which will be evident. If the machine could be programmed to know that ESSENCE, and not FOREST-, FRUIT-BEARING, is the word that needs to be retranslated, the right output, namely "example", would be obtained, because the machine could then be instructed to suspend any restrictive rule which is designed to prevent a chunk already rightly translated in Output II from being replaced by a string of synonyms. Such a rule would have to run, 'In the case of the chunk to be retranslated, reject output given by Rule X, and replace by output normally rejected by rule X. We will call this rule

Post-Editing Rule I, to show that, in this thesaurus-procedure, it cannot be automatised.

Operation 1.1 Pick out all numbers which occur more than once in Output III. Let these be called ring numbers.

result 1.1

ring number	Thesaurus item	sources of ring number
367	Vegetable	forest, fruit
161	Production	fruit, fruit, bearing
168	Productiveness	fruit, bearing
516	Meaning	bearing, essence
5	Intrinsicality	essence, essence, essence

It is worth remarking, as an incidental fact, that "The Intrinsic Meaning of the Productiveness of Vegetable Production" could stand as a sub-title, of a sort, for the whole paper.

Operation 1.2 Reorder ring numbers in order of descending frequency of occurrence. In the case of two ring numbers which occur with equal frequency, put first those which ring together most chunks. If order is then still undecided in any case, take input order.

result 1.2 5, 367, 161, 168, 516

Operation 2.1 Compare for common elements, in twos, the Thesaurus items bearing the ring numbers in the comparisons which are permitted by the lattice-relations of the chunks which are being put through the procedure, (in this case those of the invertor-lattice 60, 56, 60). In the case of any two chunks, A and B, call this comparison A B.

Order of comparisons: (i) $A \geq A$ (e.g. fruit $\cap$ fruit)

N.B. When this lattice- relation yields $\underline{a \cap a}$, *a* being not a chunk but a ring number, take output which is identical with original chunk. (example $161 \cap 161$)

(ii) A covers B

(iii) $A \geq B$

The output produced by the comparison, subject to the restrictive and permissive rules given below, is to be taken as synonymous with the chunk A in the form $A \geq A$, and with the chunk B in the case where A covers B or $A \geq B$.

Since the invertor-lattice-elements 60, 56, 60 are formed from 2-element-chains 30, 39, the following comparisons are permitted in this case.

result 2.1

lattice-relation	chunk-comparison	ring number-comparison
$A \geq A$		

≡ A+A = A	FRUIT ∩ FRUIT	367 ∩ 161
	FRUIT ∩ FRUIT	161 ∩ 168
	FRUIT ∩ FRUIT	367 ∩ 168
	BEARING ∩ BEARING	161 ∩ 168
	BEARING ∩ BEARING	161 ∩ 516
	BEARING ∩ BEARING	168 ∩ 516
	ESSENCE ∩ ESSENCE	5 ∩ 516
A covers B	Forest ∩ -Y	No comparison, as -Y has no entry
	FRUIT ∩ BEARING	161 ∩ 168
	FRUIT ∩ BEARING	161 ∩ 367
	FRUIT ∩ BEARING	161 ∩ 516
	FRUIT ∩ BEARING	168 ∩ 367
	FRUIT ∩ BEARING	168 ∩ 516
	FRUIT ∩ BEARING	367 ∩ 516
A ≥ B	FOREST ∩ ESSENCE	367 ∩ 5
≡ A ∩ B = B	FOREST ∩ ESSENCE	367 ∩ 516
(A ∩ B) ∩ C	FRUIT-BEARING ∩ ESSENCE	161 ∩ 5

≡ (A ∩ C) ∩ (B ∩ C)	FRUIT-BEARING ∩ ESSENCE	168 ∩ 5
	FRUIT-BEARING ∩ ESSENCE	161 ∩ 516
	FRUIT-BEARING ∩ ESSENCE	168 ∩ 516
	FRUIT-BEARING ∩ ESSENCE	367 ∩ 5
	FRUIT-BEARING ∩ ESSENCE	367 ∩ 516

N.B. The comparison FOREST ∩ FRUIT is prohibited, since these chunks are incomparable in the lattice. But no new comparison would result from allowing this, since all possible combinations of the five numbers already occur.

Operation 2.2 List common elements given by Thesaurus-item comparisons.

ring numbers	thesaurus-items	outputs
5 ∩ 161	Intrinsicality ∩ Production	flower; &c 22

New Comparisons Generated:

5 ∩ 22	Intrinsicality ∩ Prototype	example, specimen
161 ∩ 22	Production ∩ Prototype	pattern, prototype
5 ∩ 168	Intrinsicality ∩ Productiveness	NO OUTPUT
5 ∩ 367	Intrinsicality ∩ Vegetable	flower
5 ∩ 516	Intrinsicality ∩ Meaning	essence, example, meaning, &c 22

New Comparisons Generated:

5 ∩ 22		SEE ABOVE
516 ∩ 22	Prototype ∩ Meaning	prototype, example
161 ∩ 168	Production ∩ Productiveness	propagation, fertilisation, fructify, produce -- 168, 168 -- 161
161 ∩ 367	Production ∩ Vegetable	growth, flower
161 ∩ 516	Production ∩ Meaning	prototype &c 22

New Comparisons Generated:

161 ∩ 22		SEE ABOVE
516 ∩ 22		SEE ABOVE
168 ∩ 367	Productiveness ∩ Vegetable	NO OUTPUT
168 ∩ 516	Productiveness ∩ Meaning	NO OUTPUT
367 ∩ 516	Vegetable ∩ Meaning	NO OUTPUT

Operation 3.1 Produce synonyms for the passage required by applying outputs given under 2.2 to comparisons permitted under 2.1.

<u>synonym-outputs</u>

for FRUIT	(i) growth, flower (ii) propagation, fertilisation, fructify, produce N.B. since cross-references both from 161 to 168 and from 168 to 161 lead to permitted comparisons 161 ∩ 161 and 168 ∩ 168, apply

	2.1., i. and substitute FRUIT
for BEARING	AS ABOVE, i.e., under 2.1.i, substitute BEARING
for ESSENCE	essence, example, meaning, &c 22 example, specimen prototype, example
for FRUIT-BEARING	AS ABOVE. i.e. under 2.1., i. substitute FRUIT-BEARING
for FOREST ESSENCE	flower
for FRUIT-BEARING ESSENCE	flower, &c 22, example, specimen pattern, prototype, prototype, & 22, example, specimen, prototype, example, flower.

So far, we have used no restrictive or permissive rule except 2.1., i. If we make use of the following additional rules, to distinguish between output, we get the following final result:

Restrictive Rules

(i) 2.1., i (as above)

(ii) If a chunk of output II generates no ring number in the thesaurus, and thus generates also no comparison, replace it by itself in Output IV

By this rule, FOREST is reinserted as FOREST

(iii) If rule 2.1., i operates, reject all other output.

By this rule, FRUIT remains FRUIT, -BEARING remains -BEARING, and FRUIT-BEARING remains FRUIT BEARING

(iv) When selecting final output, take longest output first. i.e. if there is a synonym-output for FRUIT-BEARING ESSENCE, select it in preference to a synonym for FRUIT-BEARING. (This is analogous to the pidgin-dictionary matching rule, given earlier).

By using these, we remove all but the final synonyms:

Output IV: for FOREST ESSENCE forest flower

for FRUIT-BEARING ESSENCE fruit-bearing example, (3 occurrences), flower (2 occurrences), prototype (2 occurrences), specimen, (2), pattern, (1 occurrence).

N.B In this output, alternatives have been reordered in order of occurrence, and the output *&c 22* deleted. *Asterisked entries in Thesaurus*: In item 5:

item 5, row 1: ...essence, essential,...essential part,...gist,pith, core, kernel, marrow,...important part, &c, 642, *meaning, &c, 516*

row 2: principle, nature, constitution, character, type, quality; *token, example, instance, specimen &c 22;*

item 161, row 4: authorship, publication, works, opus; *result, answer, calculation; arrangement, pattern, prototype, &c 22; product, treatment*

In the case of ESSENCE, the full Thesaurus test-procedure has been given. In the other cases taken from the Italian paragraph, which follow only the results of the successive operations are shown.

Case II: SELF-PRESENT

1.2 ring numbers

such 17

problem 454

self 13, 79, 451, 486, 565, 604, 717, 836, 861, 864, 879, 880, 942, 943, 950, 953, 990

present 118, 151, 186, 457, 505, 763, 861, 894

interest 454, 455, 457, 780

(1.2. re-ordering of these in descending order of frequency of occurrence)

2.1. permitted comparisons output (153 comparisons wait for future computer)

SELF ∩

SELF

PRESENT ∩ 118 ∩ 151 Eventuality ∩ Present Time NO OUTPUT

PRESENT			
	118 ∩ 186	Presence ∩ Present Time	present
	118 ∩ 457	Present Time ∩ Attention	NO OUTPUT
	118 ∩ 763	Present Time ∩ Courtesy	present
	118 ∩ 894	Present Time ∩ Offer	NO OUTPUT
	151 ∩ 186	Eventuality ∩ Presence	NO OUTPUT
	151 ∩ 457	Eventuality ∩ Attention	concern
	151 ∩ 763	Eventuality ∩ Offer	NO OUTPUT
	151 ∩ 894	Eventuality ∩ Courtesy	NO OUTPUT
	186 ∩ 457	Presence ∩ Attention	NO OUTPUT
	186 ∩ 763	Presence ∩ Offer	NO OUTPUT
	186 ∩ 894	Presence ∩ Courtesy	NO OUTPUT
	457 ∩ 763	Attention ∩ Offer	NO OUTPUT
	457 ∩ 894	Attention ∩ Courtesy	attentive
	763 ∩ 894	Offer ∩ Courtesy	NO OUTPUT
PARTICULAR ∩ PARTICULAR	79 ∩ 151	Speciality ∩ Eventuality	NO OUTPUT
	79 ∩ 594	Description ∩ Speciality	particularise, specify
	79 ∩ 780	Speciality ∩ Property	personal
	51 ∩ 594	Eventuality ∩ Description	NO OUTPUT
	151 ∩ 780	Eventuality ∩ Property	business
	594 ∩ 780	Description ∩ Property	NO OUTPUT

INTEREST ∩ INTEREST	454 ∩ 455	Topic ∩ Curiosity	interest, &c, 461
	454 ∩ 457	Topic ∩ Attention	&c. 451
	454 ∩ 780	Topic ∩ Property	interest, business
	455 ∩ 457	Curiosity ∩ Attention	interest, attentive
	455 ∩ 457	Curiosity ∩ Property	NO OUTPUT
	457 ∩ 780	Attention ∩ Property	NO OUTPUT

By these comparisons, two new ring numbers are generated, 451, 461. These cause the ring numbers for problem now to be 454, 451, 461, and the ring numbers for interest now to be 451, 454, 455, 457, 461, 780. These additions permit the following additional comparisons of the form A ≥ A.

PROBLEM ∩ PROBLEM	454 ∩ 451	THOUGHT ∩ TOPIC	&c 461
	451 ∩ 461	THOUGHT ∩ INQUIRY	study, discuss, consider
	454 ∩ 461	TOPIC ∩ INQUIRY	&c 451; question, problem
INTEREST ∩ INTEREST	451 ∩ 454	THOUGHT ∩ TOPIC	&c 461
	451 ∩ 455	THOUGHT ∩ CURIOSITY	&c 457
	451 ∩ 457	THOUGHT ∩ ATTENTION	thought, reflection,

		consideration, interest, close study, occupy the mind, strike one as, &c 458
451 ∩ 461	THOUGHT ∩ INQUIRY	study, discuss, consider
451 ∩ 780	THOUGHT ∩ PROPERTY	NO OUTPUT
454 ∩ 455	TOPIC ∩ CURIOSITY	NO OUTPUT
454 ∩ 457	TOPIC ∩ ATTENTION	interest, &c 451
454 ∩ 461	TOPIC ∩ INQUIRY	&c 451, question, problem
454 ∩ 780	TOPIC ∩ PROPERTY	business
455 ∩ 457	CURIOSITY ∩ ATTENTION	interest, attentive
455 ∩ 461	CURIOSITY ∩ INQUIRY	prying, what's the matter?
455 ∩ 780	CURIOSITY ∩ PROPERTY	NO OUTPUT
457 ∩ 461	ATTENTION ∩ INQUIRY	&c 451
457 ∩ 780	ATTENTION ∩ PROPERTY	NO OUTPUT
461 ∩	INQUIRY ∩ PROPERTY	NO OUTPUT

At this point the detailed procedure was broken off, since it was already clear that the output of greatest frequency, among the synonyms given for *INTEREST* would be "thought, reflection, consideration, interest, close study, occupy the mind, strike one as, &c 458", namely the output of 451 ∩ 457. For the additional newly generated

ring number, 458, *Inattention*, yields only *&c 457* as output when compared with any of the others; and this output is already also given by 451 ∩ 455. Three other outputs already also include 451. Thus if the work of comparison is continued, the combination 451 ∩ 457 will increasingly recur.

It is clear that the synonyms required for idiomatic translation of SELF-PRESENT namely "strike one as, occupy the mind", will occur in the wrong position, namely as synonyms for INTEREST. Nor can this error be corrected from the lattice-program. For this, as given, allows only the comparisons SUCH ∩ PROBLEM and PARTICULAR ∩ INTEREST, neither of which will improve the synonym-output for SELF-PRESENT. The only lattice-relations which will produce the required connection are those given by the extended lattice consisting of the whole sentence, and only this after the dualising operation, has already been performed. This operation, by reversing the meets and joins of the lattice, allows SELF-PRESENT ∩ PROBLEM to occur as B-element of a 2-element chain of which PARTICULAR ∩ INTEREST occurs as A-element and thus allows 2.1., v to operate. But this intersentential lattice-programme does not exist as yet.

The final output, therefore, of this application of the procedure, is as follows:

2.2	for PROBLEM	study, discuss, consider; question, problem
	for PRESENT	present, concern, attentive
	for PARTICULAR	particularise, specify, personal, business
	for INTEREST	thought, reflection, consideration, interest, close study,

		occupy the mind, strike one as; study, discuss, consider; question, problem; business, attentive; prying, what's the matter?
	for PARTICULAR ∩ INTEREST	application, hobby, particularity, application, indicate, prove, occur, find, affair, run over, specification.

Of these last, the output of 151 ∩ 451, Eventuality ∩ Thought, prove, occur, find, is of interest, as it would be given under 2.1., v, above, since 151 is a ring number also in PRESENT.

Case III: SPROUT

1.1. **ring numbers**

with 52

sprout 35, 154, 194

reduce 144, 160

development 35, 144, 154, 194

1.2. ring numbers in order of frequency of occurrence: 35, 144, 154, 194, 52, 160

2.1 permitted comparisons

SPROUT ∩ SPROUT	35 ∩ 154	Increase ∩ Effect	production, development, grow, sprout, shoot
SPROUT			

	35 ∩ 194	Increase ∩ Expansion	increase, enlargement, augmentation, extension, growth, development, spread, swell, shoot, sprout
	154 ∩ 194	Effect ∩ Expansion	growth, development, sprout, shoot.
REDUCE ∩ REDUCE	144 ∩ 160	Conversion Weakness	reduce
DEVELOPMENT ∩ DEVELOPMENT	35 ∩ 144	Increase ∩ Conversion	growth, development, grow
	35 ∩ 154	Increase ∩ Effect	SEE ABOVE
	35 ∩ 194	Increase ∩ Expansion	SEE ABOVE
	144 ∩ 154	Conversion ∩ Effect	grow
	144 ∩ 194	Conversion ∩ Expansion	development, growth, grow
	154 ∩ 194	Effect ∩ Expansion	SEE ABOVE
REDUCE ∩ DEVELOPMENT	SEE SPROUT ∩ REDUCE ∩ DEVELOPMENT		
SPROUT ∩ REDUCE ∩ DEVELOPMENT	35 ∩ 144	Increase ∩ Conversion	growth, development, grow
	35 ∩ 154	Increase ∩ Effect	production,

		development, grow, sprout, shoot
35 ∩ 160	Weakness ∩ Increase	shoot
35 ∩ 194	Increase ∩ Expansion	increase, enlargement, augmentation, extension, growth, development, spread, swell, shoot, sprout
144 ∩ 154	Conversion ∩ Effect	grow
144 ∩ 160	Conversion ∩ Weakness	reduce
144 ∩ 194	Conversion ∩ Expansion	development, growth, grow
154 ∩ 160	Effect ∩ Weakness	bud, shoot
154 ∩ 194	Effect ∩ Expansion	growth, development, sprout, shoot

2.2. synonyms for SPROUTS, in SPROUTS AT REDUCED DEVELOPMENT;

development (5 occurrences), shoot (5 occurrences) growth, (4 occurrences) sprout (3 occurrences) production, bud, reduce, spread (1 occurrence).

Asterisked Entries in Thesaurus, for cases II and III: cross- references:

interest

concern 9

occupation

curiosity 455

etc.

sprout

grow 35

germinate 161

off-spring 167

vegetable 365, 367

expand 194

-*from* 154

item 35, *Increase*, row 2: V increase, augment, add to, enlarge; dilate &c 194; grow, wax, mount, swell, get ahead, gain strength; advance; run shoot, shoot up; rise; ascend &c, 305; sprout &c 194.

" 129, *Infant*, row 5: scion; sapling, seedling; bud, tendril, shoot, olive-branch, nestling, chicken, duckling; larva, caterpillar, chrysalis, etc.

" 160, *Weakness*, row 4: weakling; infant, &c 129.

(Delete &c 129 and insert as below:)

weakling; infant; mite, tot, little one, slip, seedling, tendril, shoot, whelp, pup, lamb; infantile, puerile, babyish, new-fledged, callow.

" 451, *Thought*, row 5: V think, reflect, reason, cognitate, excogitate, consider, deliberate; bestow thought upon, bestow consideration upon; speculate, contemplate, meditate, ponder, muse, dream, ruminate, run over; brood over; animadvert, study; bend the mind, apply the mind &c, 457; digest, discuss, hammer at, weigh prove, perpend; realise, appreciate, find; fancy, &c 515; trow.

row 9: occur; suggest itself; come into one's head, get into one's head; strike one, strike one as; be; run in one's head, etc.

" 454 *Topic*, row 1: N food for thought; mental pabulum, hobby, interest, &c 451

row 2: subject, subject-matter; theme, question topic, thesis, etc.

" 455, *Curiosity*, row 4: Adj: curious, interested, inquisitive, burning with curiosity, etc.

" 457, *Attention*, row 1: attention; mindfulness &c, adj.; intentness; thought &c, 451; advertence; observation; consideration, reflection; heed; particularity; notice, regard &c, interest, concern; circumspection, &c, 459; study, scrutiny, etc.

row 2: catch the eye, strike the eye; attract notice; catch, awaken, wake, invite, solicit, attract, claim, excite, engage, occupy, strike, arrest, fix, engross, absorb, rivet the attention, mind, thoughts; strike one, strike one as, be present to, uppermost in the mind.

item 461, *Inquiry*, row 3: sifting, calculation, analysis, specification, dissection, resolution, induction.

" 780, *Property*, row 8: money, &c 800; what one is worth; estate and effects: share- holdings, business assets, business

" 894, *Courtesy* courteous, polite, attentive, civil, mannerly, urbane, etc.

5. What is claimed for the Thesaurus-procedure is the following:

i. It is a planned procedure for producing idiomatic translation. When the translation fails, it is possible to see why.

ii. Translation-trials made by using it throw unexpected light on the principles of construction of a Thesaurus. They should, therefore, yield information which will facilitate the construction of a Thesaurus strictly compiled on statistical data for scientific MT.

iii. On this procedure, the only bi-lingual dictionaries used are word-for-word pidgin dictionaries. Nearly all the dictionary- making is done in the target language, in which the work of compiling the Thesaurus, however laborious, need only be done once, since the Thesaurus will transform the mechanical pidgin produced from all languages.

iv. The Thesaurus procedure uses previous MT results, which have established the high degree of intelligibility which can be reached by a mechanical pidgin, while at the same time keeping open the possibility of further analysing the input text.

As against this, it will be urged that for MT the whole procedure is quite impracticable, since no computer could hold a coded Thesaurus. This is true, if the Thesaurus were to be actually constructed and kept in being. The possibility exists, however, if all the items form lattices, of coding merely the chunks of the English language, together with a specification of the thesaurus positions in which each occurs. This presents a formidable coding problem; but, with modern techniques of compressed and multiple coding, not an impossible one. Once idiomatic MT is what is aimed at, a problem of comparable order would be presented by the necessity of coding, say, the two-volume concise Oxford Dictionary. Current comments on the literature, moreover, already make it clear that the commercial world is not going to be satisfied with anything short of an attempt to provide multilingual, fully idiomatic MT, since, the better the mechanical pidgin which is provided for the commercial readers' inspection, the more impatient the reader becomes with the fact that it is not wholly intelligible and correct.

7. What is a thesaurus?

Introduction

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:

First, 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 (MT) 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 MT group which is interested, so that, at last, repeatable, thesauric translations (or mistranslations) can be obtained.

I think the construction of an MT 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 or 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 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 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.

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 MT thesaurus (Mark n) survives large-scale testing on a really suitable machine.

I. Logical and mathematical account of 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 consisting 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), where a dimediate division is a binary chop. 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, listed 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.

I (b) A finite mathematical model of a Thesaurus

1. 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, 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 has correctly brought up, and which unless some special 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.1 - *Oriented partially-ordered set consisting of the dictionary entry of a word*.

INSERT FIGURE 19 HERE

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 (that 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 sublanguage-system, consisting of a sub-thesaurus.

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*.

INSERT FIGURE 20 HERE

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*:

INSERT FIGURE 21 HERE

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 $N^2 = N^1 \cup N^2$

INSERT FIGURE 22 HERE

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

a) Word Case

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

INSERT FIGURE 23 HERE

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:

INSERT FIGURE 24 HERE

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

INSERT FIGURE 25 HERE

By conflating the two partially-ordered sets, which is done by mapping the sets of contexts of the words on to the heads, - the sets being finite, as this is a finite model - we now get a single partially-ordered set with one top point and one bottom-point; that is, a partially ordered set which has a tendency to be a *lattice-like figure constructed by conflating the two oriented partially-ordered sets, given above*.

INSERT FIGURE 26 HERE

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, 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 A3/5, (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 below) 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 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 A3/5 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.

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*.

It is also crude mathematically, since the lattice A3/5, 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 A3/5 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 earlier in this chapter; 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 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 incorporating 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 A3/5. 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.

INSERT FIGURE 27 HERE

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 A3/5, 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 assistance 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.

4. 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 . . .", (see this volume). There, however, it is only applied to the semantic thesaurus, not to the language lattice as a whole .

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 (1958) 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 (Parker Rhodes & Needham, 1959).

However, the process also throws some light on the logical structure of the whole thesaurus.

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 pseudometric 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 *Evidence* 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 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[1]. 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 con-texts 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 normalisation, beyond which the distinction between the three goes bad on you; and

[1] 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.

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 2 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*).

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 let 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[2].

[Editors note: a long section has been removed here that duplicates an earlier chapter].

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 in-determinate, 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 that 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.

[2] *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.

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".

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 they say no more, instead of giving each a lengthy explanation.

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; they are not. They are distinguished 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

grouping are distinguished by sentential function. In short, the synonyms in synonym-groupings are compared to one another and distinguished form 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 MT. For basically, a word-use in context is something which you "see"...

2. 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*, which in classical form is the stop-rule theorem in Koenig's Theory of Graphs, will provide a theoretic definition of Head.

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.

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 with context; 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.

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 works 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 don 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 his own right; and even then, if inconvenient, the promotion need not be made.

The primary noticed universe remains more stable that 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

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 practise, that when points in the thesaurus-lattice are very near the top, they become so general that, by meaning practically everything, the cease to mean anything. Such points will be defined as being “above the meaning-line”. In practise, 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 noting: 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 that it is like please.

Until lately we were so impressed by this difficulty that we assumed that it was impossible, in practise, 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-make 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 archedhead 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 toe 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*.

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[3]. 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 the insights of philosophers; modern linguistics professes not to be interested, for it claims that the "word" is in no sense a fundamental notion.

So the difficulty is there, in any case. If the thesaurus is to be interlingual, there is no 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 practise, 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 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.

[3] [Editor's note: Chapter 3].

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.

Yes, but suppose what is really different as between language and language (considering no "a language" as we 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 different lattice, not a sub-lattice of a central total language lattice[4], 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 alike indeed to the weak semantic components which linguists at present use to distinguish components of a system.

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

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, 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 (1951) writes on Linguistics as applied to the particular case of the English language semantic categories, heads, descriptors - call them what 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 as an actual table from Whitehall's book (Whitehall, op cit. p.72):

Figure: The system of prepositions

RELATION	Simple Primary	Complex	Double	Group
	Transferred			

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

So, looking at this fundamental feature of linguistics from a theoretic and thesaurus-maker's point of view, we see that Haugen may have been onto a more important point that he realised when he said (1950):

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

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 Lattice. If a thesaurus or interlingua, when used on any language, produces, when tested, natural-sounding rows and lists which occur as lists in that language, then the thesaurus or interlingua has an empirical basis for that language. If the test produces arbitrary[5].

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

---

[5] 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 liost-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.

c. 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 frin ibe abitger, 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 produce 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-spokeness 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 *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 know that that head gives to the context which is being referred to in

the input text, as it is know 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 the be translated micro-glossary wise, but 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; other wise the procedure[6] will not tell you which list to use. If you 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 fill out, and if all the theoretic difficulties arising from them had been adequately encountered, this would be the end of the theoretic part of this paper.

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

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

1. 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.

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 wordpower is being increased. Fro 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 must 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 may cross-references from it is likely to have a great may cross-references to it.

f. The intersection procedure, as in "Agricola.." 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 and Needham, 1962). This was tantamount to saying that the thesaurus at present existing had a great deal less usable structure that 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 from MT 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 that 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, 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 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 sublattive concerned with the request for documents on “Linguistic Analysis and MT 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 Parker-Rhodes and Needham (1962). 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 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 efficionblyas 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 cumbrous through the great number of its terms; in an anxiety not to lose information from the system 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 mere 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 (1937). 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*[7]. 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

[7] The term "clump" was invented by Dr I.J. Good.

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 syntactic structure.

If the syntactic structure procedure is regarded as something super-added to the thesaurus, Richens' *NUDE* is an interlingua in the present sense . If the bonding[8] be disregarded, the 48 elements seem very like archeheads, and would give rise to a lattice structure with much less resolution than a whole thesaurus, but with an additional intersecting power[9].

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

[8] *NUDE* is described below in Section IV.
[9] cf. Section II, 3, above

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 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 convertibility 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 MT interlingua. This following (see overleaf) Italian-English translation trial, done on a random 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 canatteristica cui nel commercio si attribuisce assai grande importanza, dipende essenzialmente dalle sostanze coloranti naturali pres~nti nella stessa farina. Pero sul colore varie cause accessorie influiscoro e supratutto la presenza di sostanze scure estranee. La granularit' stessa della farina ha un effetto sul colere, giacche i grossi g-f'anUli proiettano un'ombra che da alla farina una sfumature bluastra.

(Genetica Agraria 1946: *1* : 38)

Out-put

The colour of the *caratteristic* flour of which very big importance is thought in connexion with commerce is conditioned naturally by the *color* natural present substance in the same flour. But different *accessor* causes and especially presence of dark *estrane* substances influence the colour. The same *granul*-ness of the flour has a effect in connexion with the colour because the big *granul*-s *proiett* a shade that gives the flour a bluish *sfumatur*.

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

NB. 2 - In the above translation, characteristic, 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.

INSERT FIGURE 28 HERE

IV. To what extent is a Thesaurus interlingual?

The extent to which any thesaurus is interlingual is, in practise, 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[10]. 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 Heights, New York in November, 1953. 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 *Intrinsically* failed; words like "real" "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 folded ones as judged by this test, which are blithely transferred from Roget thesaurus to Roget thesaurus.

[10] This information, together with other information used in this section, comes from Der Deutscher Wortschatz.

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 “Dictionaire 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*, above, it was asserted that heads, by their nature, must represent frequently noticed extra-linguistic contexts. It follows from these facts that it is contexts, no facets, 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 encyclopaedic 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, opression; 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*: *Dictionnairre Analoque*, edited by Maquet:

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

Dur d'autorite Se faire criandre. Sevir, sevices Maltraiter. Malmener. Rudoyer. Traiter de Turc a More. Parler en maitre. Parler d'autorite. Ton imperatif. Ne pas badiner. Montrer les dents. Cassant. Rembarrer., - Discipline. Main de fer. Inflex ible. Rigide. Severe. Strict Tenace. Rigoreux. Exigeant. - Terrible. Tyrannique. Brutal Despotique. - Rebarbatif. Pas commode. Grandeur. Menacant Cerbere. Intimider.

3. *German*: *Deutscher Wortschatz*, Head 739 *Strenge*.

Harte. Unerbittlichkeit. Unerschutterlichkeit. Harterzigkeit. Herzenshartigkeit. Grausamkeit. Rucksichtslosigkeit. Gemeinheit. Unduldsamkeit. (Intoleranz). Rechthaberei. Unnachsichtigkeit.

From a comparitive 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 a 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, any 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 you own set, uncover the column again, and compare[11].

[11] 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 idosyncracy; though the frequen appearance of such phrases in NUDE entries also suggests otherwise; if the tendency to use verbal phrases for row definition is a

ROGET'S ROWS	DISCURSIVE DESCRIPTION OF ROW
*white*ness	people think of the abstract notion of WHITENESS; a colour
snow, paper, chalk, milk, lily, ivory; white lead, chinese *white*, white-wash, whitening	white concrete objects, both solid and liquid
render *white*, blanch, *white*-wash, silver, frost	the action of causing something to become white
white; milky, milk-*white*, snow-*white*, snowy, candid	people see objects having a white appearance
white as a sheet; white as the driven snow	concrete whiteness of colour being used to symbolise mental states of FEAR, INNOCENCE
vision, sight, optics, eye-sight	the faculty of seeing
visual organ, organ of vision, eye	the part of the body with which a man sees
eye-ball, retina. pupil, iris, cornea, *white*	list of parts of the eye
abject fear, funk	people exhibiting this
white feather, faint-heart, milk-sop, *white* liver, cur, craven	picturesque statements of the appearance and physiology of people exhibiting COWARDICE

natural one, then the criticism that a “thesaurus” is a system consisting only of nouns (G.W. King) is unfounded.

faint-hearted, chicken-hearted; yellow, *white*-livered	people abusing their fellows in concrete terms for exhibiting COWARDICE
etc.	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.) [….]