APPENDIX III

(Workpaper referred to in course of Text)

"AGRICOLA INCURVO TERRAM DIMOVIT ARATRO"

(Virgil, "Georgics")

First Stage Translation

into English

with the aid of

Roget's Thesaurus

Copyright:
The Cambridge Language Research Unit
20 Millington Road
Cambridge, England

Carried out in
November 1957
Margaret Masterman
R.M. Needham
K. Sparck Jonea
Bryan Mayoh
OUTLINE OF A THESAURUS- USING TRANSLATION PROGRAMME

(Latin to English)
Using Roget's Thesaurus, Penguin Edition

The essential feature of this programme, is the use of a thesaurus as an interlingua: the translation operations are carried out on a head* language into which the input text is transformed and from which an output is obtained. These operations are of three kinds: semantic, syntactic and grammatical.

The general arrangement of the programme is as follows:

I. Dictionary matching: the chunks of the input language are matched with the entries in a Latin-Interlingual Dictionary giving the raw material of head language; this consists of heads representing the semantic, syntactic and grammatical elements of the input.

II. Operations on the semantic heads: these give a first stage translation.

III. Operations on the syntactic heads: giving a syntactically complete, though unparsed, translation.

IV. Operations on the grammatical heads: giving a parsed and correctly ordered output.

V. Cleaning up operations: The output is "trimmed" by e.g. insertion of capital letters, removal of repetitions like "farmer-er".

This programme is based partly on an interlingual translation programme by R.H. Richens published in July, 1957 as a workpaper of the C.L.R.U. entitled The Thirteen Steps; partly on a thesaurus-using translation procedure by Margaret Masterman, from a paper entitled Potentialities of a Mechanical Thesaurus, read at the 2nd International Conference on Machine Translation (M.I.T. Octo. 17th, 1956);

* The notion of "heads" is taken from the concepts of topics under which Roget classified words in his thesaurus.
and partly on a library retrieval procedure making use of a thesaurus by T. Joyce and R.M. Needham described in a C.L.R.U. workpaper entitled The Thesaurus Approach to Information Retrieval.

Only Stage II of the procedure is given in detail here.

**INFORMATION OBTAINED FROM STAGE I**

The Latin sentence to be translated was chunked as follows:

AGRI-COL-A INCURV-O TERR-AM DI-MOV-IT AR-ATRO

A number of these generated syntactic heads only. Those with semantic head entries are AGRI- -COL- IN— CURV- TERR- DI— -MOV- AR-.

The interlingual dictionary entries, for each chunk were constructed by a transformation into thesaurus heads of the information given in Lewis' "Latin Dictionary for Schools" for all words containing the chunk in question. This can be followed by comparing the semantic head sets and the dictionary entries taken from Lewis' Dictionary given overleaf.
SAMPLE HEAD SETS CORRESPONDING WITH LEWIS' DICTIONARY ENTRIES FROM WHICH THEY WERE MADE.

AGRI-
181 REGION
189 ABODE
371 AGRICULTURE
780 PROPERTY
SAMPLE HEAD SETS CORRESPONDING WITH LEWIS' DICTIONARY ENTRIES FROM WHICH THEY WERE MADE.

**TERR- (1)**

181 REGION
211 BASE
318 WORLD
342 LAND
673 PREPARATION

**TERR- (2)**

668 WARNING
669 ALARM
378 PAIN
860 FEAR
887 BLUSTERER
### INTERLINGUAL DICTIONARY

<table>
<thead>
<tr>
<th>AGRI-</th>
<th>-COL-</th>
<th>-IN-</th>
</tr>
</thead>
<tbody>
<tr>
<td>181 REGION</td>
<td>188 INHABITANT</td>
<td>54 COMPOSITION</td>
</tr>
<tr>
<td>189 ABODE</td>
<td>186 PRESENCE</td>
<td>176 TENDENCY</td>
</tr>
<tr>
<td>371 AGRICULTURE</td>
<td>758 CONSIGNEE</td>
<td>221 INTERIORITY</td>
</tr>
<tr>
<td>780 PROPERTY</td>
<td>371 AGRICULTURE</td>
<td>232 ENCLOSURE</td>
</tr>
<tr>
<td></td>
<td>342 LAND</td>
<td>247 CONVOLUTION</td>
</tr>
<tr>
<td></td>
<td>876 COMMONALTY</td>
<td>264 MOTION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>259 FURROW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>278 DIRECTION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>286 APPROACH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>294 INGRESS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300 INSERTION</td>
</tr>
</tbody>
</table>

**-CURV-**

244 ANGULARITY
245 CURVATURE
279 DEVIATION

<table>
<thead>
<tr>
<th>TERR- (1)</th>
<th>TERR- (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>181 REGION</td>
<td>668 WARNING</td>
</tr>
<tr>
<td>211 BASE</td>
<td>669 ALARM</td>
</tr>
<tr>
<td>318 WORLD</td>
<td>378 PAIN</td>
</tr>
<tr>
<td>342 LAND</td>
<td>860 FEAR</td>
</tr>
<tr>
<td>673 PREPARATION</td>
<td>887 BLUSTERER</td>
</tr>
</tbody>
</table>

**DI-**

| 44 DISJUNCTION | 371 AGRICULTURE |
| 49 DECOMPOSITION | 61 DERANGED |
| 91 BISECTION | 140 CHANGE |
|               | 264 MOTION |
|               | 673 PREPARATION |
|               | 615 MOTIVE |
|               | 824 EXCITATION |
|               | 49 DECOMPOSITION |
|               | 44 DISJUNCTION |
|               | 259 FURROW |

**AR- (1)**

| 371 AGRICULTURE | 340 DRYNESS | 1000 TEMPLE |
| 259 FURROW | 384 CALEFACTION | 903 MARRIAGE |
| 248 CONVOLUTION |          |              |
| 876 COMMONALTY |          |              |
DISCURSIVE DESCRIPTION OF THE SET OF OPERATIONS USED ON SEMANTIC HEADS; THAT IS, OF STAGE II OF THE TRANSLATION PROCEDURE

A. Elimination of unwanted heads by intersection:
   i) Standard Procedure:
      It is assumed that those semantic concepts relevant to
      the sentence to be translated will occur repeatedly
      (i.e. at least more than once). Selection of the heads
      representing those concepts could therefore be obtained
      by an intersection procedure as follows:
      Each member of the head set representing a chunk is
      matched in turn with all other heads occurring for other
      chunks in the sentence. Only those occurring twice or
      more are retained.
   ii) Removal of Puns:
      This procedure should eliminate puns: a chunk such as
      TERR- has two completely different sets of uses, repre-
      sented by completely different sets of heads, only one
      of which is relevant in a particular context. The un-
      wanted heads will probably fail to occur elsewhere in
      the sentence so that only the relevant heads represent-
      ing the appropriate chunk in question are retained.
   iii) Scale of Relevance Procedure:
      It may happen that all members of the head set for a
      particular chunk fail to intersect. In this case, we try
      to find heads in the rest of the sentence which are close-
      ly related to the heads in this set. For the present test
      heads which are within the same bracket in the Table of
      Contents in Roget's Thesaurus, are regarded as closely
      related. The procedure is as follows: all the heads
      occurring in the same bracket of the Table of Contents
      as those already given for a non-intersecting chunk are
      introduced; from a practical point of view they are re-
      garded as representing a new chunk in the sentence. The
      intersection procedure can be carried out as before. If
      unsuccessful, the manoeuvre can be repeated using bigger
      brackets in the Table of Contents. It should be noted
      that the introduction of these new head sets may increase
      the number of intersections for other chunks in the sent-
      ence. After the intersection has been carried out, the
heads retained for the new chunk are amalgamated with those of the chunk which generated it.

We now have for each unit of head language a group of heads which have shown themselves to be relevant to the subject under discussion. Thus for -MOV- we have:

- AGRICULTURE
- PREPARATION
- DECOMPOSITION
- DISJUNCTION
- FURROW
LIST OF HEADS IN BRACKETS 'SPECIAL FORM' AND 'MOTION, WITH
REFERENCE TO DIRECTION' REQUIRED FOR EXTENDED INTERSECTION
PROCEDURE

SPECIAL FORM
244 ANGULARITY
245 CURVATURE 246 STRAIGHTNESS
247 CIRCULARITY 248 CONVOLUTION
249 ROTUNDITY

MOTION WITH REFERENCE TO DIRECTION
278 DIRECTION 279 DEVIATION
280 PRECESSION 281 SEQUENCE
282 PROGRESSION 283 REGRESSION
284 PROPULSION 285 TRACTION
286 APPROACH 287 RECESSION
288 ATTRACTION 289 REPULSION
290 CONVERGENCE 291 DIVERGENCE
292 ARRIVAL 293 DEPARTURE
294 INGRESS 295 EGRESS
296 RECEPTION 297 EJECTION
298 FOOD 299 EXCRETION
300 INSERTION 301 EXTRACTION
302 PASSAGE
303 OVERSTEP 304 SHORTCOMING
305 ASCENT 306 DESCENT
307 ELEVATION 308 DEPRESSION
309 LEAP 310 PLUNGE
311 CIRCUITION
312 ROTATION 313 EVOLUTION
314 OSCILLATION
315 AGITATION
N.B. The thesaurus has been expanded so as to allow of the insertion of a set of curve-producing tools (of which Roget takes cognisance of only one member, corkscrew) under CONVOLUTION. Roget classifies a plough-share as a cutting-edge; but not as a device for turning over the sod. In fact, ploughs, c.q.r. anchors, etc. are less convoluted than horns, serpents and corkscrews, but more convoluted than horse-shoes, crooks or sickles; and therefore constitute an intermediate head. Lacking courage to construct this, we have classed them under CONVOLUTION.

The introduction of SPECIAL FORM is due to the failure of -CURV- to intersect with any of the other words. We therefore introduce, as a new chunk all the other heads in the bracket titled 'SPECIAL FORM' which includes ANGULARITY and "CURVATURE" given by -CURV-. We can then obtain our intersections. The bracket titled MOTION, with reference to DIRECTION was also introduced as it includes DEVIATION which is also given by -CURV-. This did not however result in any intersections and was therefore eliminated.
b) Selection of correct output word from the select head-sets representing each chunk.

Here the actual translation from head language to output language is made. (As the output language is English, the interlingual thesaurus, Roget, can still be used. This need not necessarily be the case.) The procedure is as follows: the contents of each head retained for a chunk are compared in turn with those of all other heads retained for that chunk. Any word which occurs more than twice is retained as output. This output constitutes a first stage semantic translation of the text. (It is obvious that difficulties may occur either if no intersection is obtained, or if there is only one head retained for a word.)
OUTPUT OF SET OF TRANSLATION-INTERSECTIONS TO OBTAIN WORDS OF OUTPUT TEXT

(An example in full is given on the next page of a translation between two heads.)
(The notation used below is to be interpreted as follows:
A ^ B = C ----- C is to be interpreted, "When the list of synonyms given by Roget under the Head A is compared with the list of synonyms given by Roget under Head B, the series of words C1----Cn, which we will call the output, will be found to occur in both lists of synonyms.) The output of these intersections should be referred to any words having the two heads concerned. E. g. Agriculture ^ Furrow relates both to -MOV- and AR-.

1. Agriculture ^ Region = etc, 189
2. Agriculture ^ Land = Farmer
3. Agriculture ^ Commonalty = Ploughman, tiller of the soil, rustic
4. Agriculture ^ Furrow = Plough
5. Agriculture ^ Convolution = No output
6. Agriculture ^ Preparation = Till, cultivate the soil
7. Agriculture ^ Decomposition = No output
8. Agriculture ^ Disjunction = No output
9. Agriculture ^ Motion = No output
10. Land ^ Preparation = No output
11. Land ^ Commonalty = No output
12. Land ^ Region = Ground, soil
13. Region ^ Preparation = No output
14. Furrow ^ Convolution = No output
15. Furrow ^ Commonalty = No output
16. Angularity ^ Curvature = Bend, etc. 217
17. Angularity ^ Convolution = No output
18. Curvature ^ Convolution = Curl
19. Convolution ^ Commonalty = No output
20. Disjunction ^ Decomposition = Disperse, etc. 73, break up
21. Convolution ^ Motion = No output
22. Disjunction ^ Preparation = No output
23. Disjunction ^ Furrow = No output
24. Disjunction ^ Motion = No output
25. Decomposition ^ Preparation = No output
26. Decomposition ^ Furrow = No output
27. Decomposition ^ Motion = No output
28. Preparation ^ Furrow = No output
29. Preparation ^ Motion = cultivator, cultivate
30. Furrow ^ Motion = No output

If two Heads have a common cross reference, this head should be included in the intersection procedure. We now bring down:

73 Dispersion
189 Abode
217 Obliquity

We then reinsert ABODE in the head-set of AGRI-, (where it once belonged). OBLIQUITY we insert in the Head-sets of -CURV- and SPECIAL FORM, which both contain ANGULARITY and CURVATURE as members); and we insert DISPERSION as an extra head in the Head-sets of DI- and -MOV-, both of which have both DISJUNCTION and DECOMPOSITION as members. We then perform a further set of intersections as follows:

31. Agriculture ^ Abode = farm
32. Region ^ Abode = etc. 232
33. Angularity ^ Obliquity = incline, bend, crook, crooked
34. Curvature ^ Obliquity = bend, crook, etc, 245
35. Convolution ^ Obliquity = twist
37 Decomposition  \^ Dispersion = No output
38. Agriculture \^ Dispersion = sow
39. Preparation \^ Dispersion = No output
40. Farrow \^ Dispersion = No output

We now bring down:

44 Disjunction
232 Enclosure
245 Curvature

of which we retain only ENCLOSURE, (under AGRI-, since both the others already exist under the relevant heads.)

We thus get the further set of intersections:

41. Agriculture \^ Enclosure = No output
42. Abode \^ Enclosure = No output
43. Region \^ Enclosure = No output
EXAMPLE OF METHOD OF TRANSLATION INTERSECTION

259  Farrow - N. farrow, groove, rut, scratch, streak, stria, crack, score, incision, slit; chamfer, fluting, channel, gutter, trench, ditch, dike, dyke, moat, fosse, trough, kennel; ravine, etc. 198.

V, furrow etc, n; flute, groove, carve, corrugate, plough, incise, chase, enchase, grave, etch, bite in, cross-hatch.

Adj. furrowed etc, v; ribbed, striated, fluted; corduroy,

371  Agriculture - N, agriculture, cultivation, husbandry, farming, agronomy; georgics; tillage, tilth, gardening, vintage; hort-, arbor-, silv-, vit-, floriculture; intensive culture; landscape gardening; forestry, afforestation,

husbandman, horticulturist, gardener, florist; agriculturist; yeoman, farmer, cultivator, tiller of the soil, ploughman, sower, reaper; woodcutter, backwoodsman, forester; vine grower, vintager.

field, meadow, garden; botanic-, winter-, ornamental-, flower-, kitchen-, market-, hop- garden; nursery; green-, hot-, glass-, house; conservatory, cucumber-, cold-frame, cloche, bed, border; lawn; park etc, 840; parterre, shrubbery, plantation, avenue, arboretum, pinery, orchard; vineyard, vinery orangery; farm etc. 189.

V. cultivate; till; farm, garden; sow, plant; reap, mow, cut, crop etc. 789; manure, dig, delve, dibble, hoe, plough. harrow, rake, week, lop and top, force, transplant, thin out; bed out, prune, graft.

Adj. agricultural, -arian, arable; rural, rustic, country, bucolic; horticultural.

The procedure consists in comparing the above sections word by word, from which it will be seen that the common output is plough.

WARNING: The use of hyphens in Roget's Thesaurus is ambiguous, since the constituent words of a hyphenated sequence of words, e.g. set - shoot - up, are not repeated within the same head, even though set, and set up can be synonyms of one another.

In this matter the person operating the thesaurus must use his own judgment.
**SEMANTIC TRANSLATION OF THE TEXT**

(that is, translation with the syntax unresolved with DI- and -MOV- combined and IN- and -CURV- combined)

<table>
<thead>
<tr>
<th>AGRI-</th>
<th>-COL-.</th>
<th>INCURV-</th>
<th>TERR-</th>
</tr>
</thead>
<tbody>
<tr>
<td>farm</td>
<td>farmer</td>
<td>bend</td>
<td>ground, soil</td>
</tr>
<tr>
<td>pen</td>
<td>ploughman, tiller of the soil, rustic</td>
<td>incline, bend crook, crooked bend, crook, twist</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DIMOV-</th>
<th>-AR-</th>
</tr>
</thead>
<tbody>
<tr>
<td>plough</td>
<td>ploughman tiller of the soil</td>
</tr>
<tr>
<td>cultivation</td>
<td>rustic</td>
</tr>
<tr>
<td>till, cultivate the soil</td>
<td>plough</td>
</tr>
<tr>
<td>cultivate</td>
<td></td>
</tr>
<tr>
<td>disperse, break up</td>
<td></td>
</tr>
<tr>
<td>sow</td>
<td></td>
</tr>
</tbody>
</table>

N.B, There is no output for IN-. This in fact reflects the somewhat redundant character it has.

The syntactical and grammatical operations must now be carried out to choose between these alternatives, to re-order the whole sentence and to introduce the additional elements which are necessary to make - the output a correct English sentence.
The head set of the new "chunk" shown as a lattice so that the procedure for applying the scale of relevance procedure may be made precise.

NOTE. It can be seen that the use of the bracket group of heads as described in the scale of relevance procedure, can be looked at from another point of view as utilising the lattice property of language. Made more precise, the procedure is: compare each head in the head-set of the non-intersecting chunk (in this case -CURV-) with the table of contents, (this last being arranged as a lattice). If, to find a common idea between any two heads in the head set not more than two steps need be taken up the lattice, bring this common idea down as a new chunk in the input text, this new chunk being inserted after the original non-intersecting chunk. (Thus SPECIAL FORM, the new chunk, will be inserted after -CURV-).

See whether any of the heads in the head set of the new chunk intersects with any head of any of the head sets in the chunks of the input text.

If an intersection is obtained, amalgamate the head sets of SPECIAL FORM and -CURV- to form a single head-set.

If no intersection is obtained, extend the procedure to bring down the second scale of relevance (i.e. in this case, bring down all the heads given in Roget's table of contents under (GENERAL, SPECIAL and SUPERFICIAL FORM and try again for an intersection, as above.

If it is still the case that no intersection is obtained, the chunk -CURV-, (or more probably the whole word INCURVO) becomes an untranslateable word of head language, - as it might be a foreign word - and is carried through complete into the English output, all heads being given in the English output text.