"AGRICOLA INCURVO TERRAM DIMOVIT ARATRO"

(Virgil, "Georgics")

First stage translation into English
with the aid of
Roget's Thesaurus

M. Masterman, R.M. Needham, K. Sparck Jones, B. Mayoh

Cambridge Language Research Unit, Cambridge, England

Report (ML84) ML92, November 1957

Reprinted, with a new Introduction by K. Sparck Jones

Computer Laboratory, University of Cambridge, Cambridge, England
April 1986

Introductory Note, 1986

Karen Sparck Jones

Much of the machine translation (MT) research of the fifties and sixties focussed on syntax; however some groups, notably the Cambridge Language Research Unit (CLRU), argued that semantics was much more important. The CLRU addressed the problem of lexical disambiguation, and advocated the use of a thesaurus as a means of characterising word meanings, in part because the structure of a thesaurus naturally supports procedures for determining the senses of words or, complementarily, for finding words for meanings. The assumption is that text has to be repetitive to be comprehensible so, in the simplest case in disambiguation, if a word's senses are characterised by several thesaurus classes, or heads, the relevant one will be selected because it is repeated in the list for some other text word. In text production, the fact that two heads share a word suggests that this is the right one; for translation this mechanism could provide a means of selecting appropriate target language equivalents for source words.

The thesaurus could thus be seen, for translation, as constituting an interlingua; and as it appeared it could be formally modelled as a lattice, procedures using it could be formally specified as lattice operations. It was further argued that syntax, and grammar, could be approached through the thesaurus, though this was never worked out in detail. In particular, the relation between syntax and semantics in text processing was never properly specified, though one strength of the way a thesaurus was used for disambiguation was that its application was not narrowly constrained, as it was later by Katz and Fodor, by syntactic structure. But equally, the experiments done were very simple, so the need to relate syntactic and grammatical information to semantic information in processing was underestimated. Actual tests on sense selection in the translation context tended to retain input word order in the initial output, for hypothesised rearrangement for the final output.

The experiment described in this Workpaper is part of a series carried out by the CLRU in the late fifties: Latin was chosen as as the input language as the only one apart from English common to all members of the CLRU. The experiments could not be carried out automatically, as the CLRU had no computer, but were done 'mechanically', i.e. by working with paper lists in the style required for the procedures using punched card apparatus then being devised at the Unit. The essence of the experiment described here was to select the appropriate senses of words, or rather of their semantically-significant morphological components, referred to as chunks, by selecting those heads in each chunk's list which were shared with some other chunk; and then to obtain the corresponding English chunk (in fact word) as any item common to the heads in each chunk's list. The procedure included strategies for dealing with failures to obtain any common

heads or words. (I have detected minor errors in the text, not affecting the results, and have simply corrected them; I have not attempted to replicate the test. I have retained the original spelling of “programme” as historically appropriate.)

This test, like the other CLRU ones, was a very limited one. But the experiments the CLRU did were tests of well-defined procedures. The idea of using a thesaurus was a very attractive one, and the CLRU’s ideas on the semantic aspects of natural language processing were known to other research workers at the time. But they were very inadequately reported. I have reproduced this Workpaper to make the CLRU’s ideas somewhat better known, as they deserve to be, than they are.

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 [1] 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 The Potentialities of a Mechanical Thesaurus, read at the 2nd International Conference on Machine Translation (M.I.T. Oct. 17th 1956); and partly on a library retrieval procedure making use of a thesaurus devised 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

[l] The notion of "heads" is taken from the concepts or topics under which Roget classified words in his thesaurus.

The Latin sentence to be translated was chunked as follows:

AGRI-COL-A IN-CURV-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 below.

SAMPLE HEAD SET CORRESPONDING WITH LEWIS' DICTIONARY ENTRY FROM WHICH IT WAS MADE

AGRI-

181 REGION
189 ABODE
371 AGRICULTURE
780 PROPERTY

AGER, GRI, ... I. In a restricted sense, improved or productive land, a field, farm, estate, arable land, pasture etc: [quotes]. II. In an extended sense. A. Territory, district, domain, the soil belonging to a community. [quotes]. B. the fields, the open country, the country: [quotes]. C. Poet. plain, valley, champaign: [quotes].

AGRICOLA, AE, ... I. Prop. a husbandman, agriculturer, ploughman, farmer, peasant: [quotes]. II. Praegn. a rustic, boor, clown: [quotes].

(I have omitted the quotes; the head sets shown for "terr-" in the original report are not given here as the supporting illustrations reproduced from Lewis are illegible. KSJ)

SEMANTIC HEAD SETS OF THE INPUT TEXT GIVEN BY THE INTERLINGUAL DICTIONARY

AGRI-		-COL-		-IN-	
181	REGION	188	INHABITANT	54	COMPOSITION
189	ABODE	186	PRESENCE	176	TENDENCY
371	AGRICULTURE	758	CONSIGNEE	221	INTERIORITY
780	PROPERTY	371	AGRICULTURE	232	ENCLOSURE

		342	LAND	247	CONVOLUTION
		876	COMMONALTY	264	MOTION
				259	FURROW
				278	DIRECTION
				286	APPROACH
				294	INGRESS
				300	INSERTION
	-CURV-		TERR- (1)		TERR- (2)
244	ANGULARITY	181	REGION	668	WARNING
245	CURVATURE	211	BASE	669	ALARM
279	DEVIATION	318	WORLD	378	PAIN
		342	LAND	860	FEAR
		673	PREPARATION	887	BLUSTERER
	DI-		-MOV-		AR- (1)
44	DISJUNCTION	371	AGRICULTURE	371	AGRICULTURE
49	DECOMPOSITION	61	DERANGED	259	FURROW
91	BISECTION	140	CHANGE	248	CONVOLUTION
		264	MOTION	876	COMMONALTY
		673	PREPARATION		
		615	MOTIVE		
		824	EXCITATION		
		49	DECOMPOSITION		
		44	DISJUNCTION		
		259	FURROW		
	AR- (2)		AR- (3)		
340	DRYNESS	1000	TEMPLE		
384	CALEFACTION	903	MARRIAGE		

DISCURSIVE DESCRIPTION OF THE SET OF OPERATIONS USED ON SEMANTIC HEADS; THAT IS. OF STAGE II OF THE TRANSLATION PROCEDURE

A. Elimination of unwanted beads 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 heads, only one of which is relevant in a particular context. The unwanted heads will probably fail to occur elsewhere in the sentence so that only the relevant heads representing 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 closely 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(s) of the Table of Contents as those already given for a non-intersecting chunk are introduced; from a practical point of view they are regarded 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 sentence. 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
MOTION

LIST OF HEADS IN BRACKETS "SPECIAL FORM" AND "MOTION WITH REFERENCE TO DIRECTION" REQUIRED FOR EXTENDED TRANSLATION 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

THE SETS OF HEADS IN HEAD LANGUAGE AFTER
THE NON-INTERSECTING HEADS HAVE BEEN ELIMINATED

AGRI-
REGION
AGRICULTURE

-COL-
AGRICULTURE
LAND
COMMONALTY

IN-
CONVOLUTION
FURROW
MOTION

-CURV-
ANGULARITY
CURVATURE

SPECIAL FORM
ANGULARITY
CURVATURE
CONVOLUTION

TERR-
REGION
LAND
PREPARATION

DI-
DISJUNCTION
DECOMPOSITION

CONVOLUTION

-MOV-
AGRICULTURE
PREPARATION
DECOMPOSITION

AR-
AGRICULTURE
FURROW

DISJUNCTION COMMONALTY
FURROW
MOTION

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, cqr 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 later 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—C2, 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	=	fanner
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	=	cultivation, 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
36.	DISJUNCTION	^	DISPERSION	=	disperse, etc. 44
37.	DECOMPOSITION	^	DISPERSION	=	no output
38.	AGRICULTURE	^	DISPERSION	=	sow
39.	PREPARATION	^	DISPERSION	=	no output
40.	FURROW	^	DISPERSION	=	no output
41.	MOTION	^	DISPERSION	=	no output

We now bring down

44	DISJUNCTION
232	ENCLOSURE
254	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:

42.	AGRICULTURE	^	ENCLOSURE	=	no output
43.	ABODE	^	ENCLOSURE	=	no output
44.	REGION	^	ENCLOSURE	=	no output

EXAMPLE OF METHOD OF TRANSLATION-INTERSECTION

259 Furrow - N. furrow, 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-, flor-iculture; intensive culture; landscape gardening; forestry, afforestation.
husbandman, horticulturist, gardener, florist; agriculturalist; 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 with IN- and -CURV- combined)

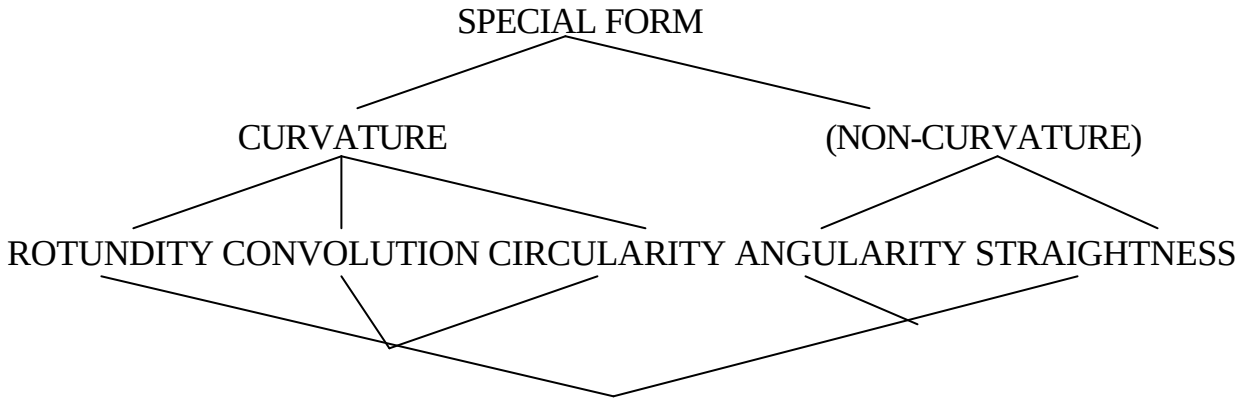
AGRI-	-COL-	INCURV-
farm	farmer ploughman, tiller of the soil, rustic	bend incline,bend, crook, crooked bend, crook twist
TERR-	DIMOV-	AR-

ground, soil	plough till, cultivate the soil cultivation, cultivate disperse, break up sow	ploughman, tiller of the soil, rustic plough

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 reorder the whole sentence and to introduce the additional elements which are necessary to make the output a correct sentence.

C. The head set of the new "chunk" shown as a lattice so that the procedure for applying the scale of relevance 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.