CHAPTER 5: Groups and projects in the United Kingdom and Western Europe (1954-66)

5. 1: Birkbeck College, London (1955-1959)

For some time after its initial formulation, the only test of Booth's and Richens' proposed mechanical dictionary for MT (ch.2.2 above) had been on punched card machinery. The basic problem, as Booth pointed out on numerous occasions, was the grossly inadequate storage capacity of machines designed for numerical calculation. Not until 1953 was Booth able to give some figures for running a word-for-word MT using Richens' stem and affix method (Booth 1953). The test was run on the APEXC computer designed by Booth and built at Birkbeck College in the University of London. The program was slow: the fastest time achieved was for a 1000 word passage using the APEXC and tabulator output (30 minutes); when using teletype output the time needed was 2 hrs 15 mins! In addition, of course, Booth had to admit the output was "inelegant" although he contended it could be "easily understood by a person expert in the subject of the paper." Booth was one of the British pioneers in computer design and construction, and his subsequent work on MT at Birkbeck was part of the research of the Electronic Computer Laboratory, which was primarily concerned with developing computational techniques and improving computer hardware (Booth 1980)[1].

Booth had returned from the 1952 MIT conference convinced, like Dostert (ch.4.3), in the urgency of constructing prototype MT systems. In September 1955 at a conference on 'Information Theory', Booth (1956a) contended that it was of "great importance to achieve some results in the mechanical translation of language in the near future, otherwise the whole subject is likely to fall into disrepute." Booth had, therefore, decided on a small-scale experiment with a "language which is sufficiently restricted to make possible its running on an existing machine" and chose as a "worthwhile project... the conversion of Standard English into Braille" in the belief that "problems which occur in translation to contracted Braille occur also in the translation of real language". This was the motive behind the work of John P.Cleave on Braille transcription (cf. Booth et al. 1958).

By this time Booth had developed his method of fast dictionary lookup, the 'binary division' technique, which he usually called the 'logarithmic method' (Booth 1955a). During 1955 the Birkbeck laboratory obtained the sponsorship of the Nuffield Foundation for "a modest programme... to render scientific French into acceptable English" (Booth 1956) Booth had selected French, firstly, in order to "avoid competition with American projects" and secondly, because French was seen as "an easy starting point". It was evidently in connection with this project that a "microglossary and micro-grammar in meteorological French" had been constructed in a joint project with a group working under Professor J.R. Firth, professor of linguistics at what is now the School of Oriental and African Studies of the University of London.

The Birkbeck approach was characterised by Booth (1965) thus: "our program started off from zero on the assumption that we could do word-for-word translation (which of course we can't) and then worked its way up through an increasing list of complications...", i.e. the empirical cyclic method (ch.4.4 and 8.2). As at many other places at the time, there was considerable ignorance, if not naivety, over the linguistic problems of translation. Typical are the remarks of Cleave & Zacharov (1955) on the generation of target language equivalents: "This involves two things. First, adding the correct word-endings. This is a simple procedure according to the rules of the 'target' language since the sentence structure and grammatical function of each foreign language word have already been determined. Secondly, arranging the 'target' language word equivalents in the conventional word order of that language. This is an entirely mechanical operation...". Judgment should be tempered, however, in recognition of the formidable computational difficulties involved;

[1] For his memoirs see A.D.Booth and K.H.V.Booth: 'The beginnings of MT', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 253-261.

the APEXC itself had only recently become reasonably reliable in operation (Booth et al., 1958) and programming techniques had advanced hardly beyond direct machine coding (cf. the chapter on MT programming in K. Booth, 1958).

The French-English MT project was the work of Leonard Brandwood. Within six months (Booth 1958), he had developed a program for French analysis. This was conceived primarily as the identification of stems and endings, described by Brandwood (1956) in considerable detail. Syntax was considered unproblematic since French "word order is more or less identical with that of English". A program for rearrangement was written but could not be implemented because of the APEXC's limited storage capacity (Cleave 1957). For the same reason, the system could be tested only on a very small 'micro-glossary' of, it would appear, no more than about 1000 words. The results were encouraging. An example translation of a mathematics text, "undistinguishable from that produced by many human operators", was given by Booth (1958):

> We demonstrate in this chapter the fundamental theorem of M. Poincaré, after having studied the integrals of a system of differential equations, considered as functions of initial values

Booth was so encouraged that shortly afterwards he felt able to write: "We are of the opinion that most of the problems which attend the translation of French into English by machine have now been solved" (Booth 1956). All that seemed to be necessary was a larger computer. The team were now "examining the application of the computing machine to language in general" and "work is now proceeding on the analysis of the German language..." using the "invaluable analysis of Oswald and Fletcher" (cf.ch.2.4)

The report of this research (Booth et al. 1958) reveals the care taken by the team to gather information about German grammar which they believed could be formalized in a way suitable for MT analysis. Some of this work by Brandwood (1958) on the German relative clauses and prepositional phrases foreshadowed future problems for MT. In order to account fully for the sequencing of noun phrases within relative clauses Brandwood concluded that some semantic information must be included in dictionary entries. For example in:

> Allerdings wird die Wirkung dieser Felder auf Elektronen, welche sie zu verschiedenen Zeiten durchlaufen, verschieden sein.
>
> (To be sure, the effect of these fields on electrons which traverse them at different times will be different.)

either *sie* (referring to *Felder*) or *welche* (referring to *Elektronen*) could be subject of *durchlaufen*. The only way of knowing which analysis should be made is with information that 'electrons' can 'traverse' 'fields' but not vice versa. "A dictionary for the machine must be compiled which classifies words and indicates not only which ones can be constructed together but also in what way." Similarly with prepositional phrases. As Brandwood put it: "In translating...

> Wir haben darauf hingewiesen, dass die Laplacesche Gleichung für die elektronenoptischen Felder gegenüber den lichtoptischen Medien eine Einschränkung bedeuten.

the English word order varies according to whether neither, one, or both prepositional phrases are interpreted as dependent on the preceding noun *equation*." Brandwood believed that certain standard sequencing rules might be possible, but recognising the ad hocness of such solutions suggested it would be "more satisfactory to have a system of word classification on the lines suggested (for) the relative pronoun", i.e. indicating semantic features.

It is apparent that the Birkbeck team was not able to program the German-English system, primarily because of insufficient storage capacity in the computers available. In effect, active work on MT at Birkbeck had died away by 1959, but already by this time there was a strong interest in applications of the computer to other linguistic processes (Booth et al. 1958), and in later years, increasing attention was paid to programs for stylistic analysis, for the analysis of text statistics and for the production of concordances (cf. Levison 1962)

Research on computational linguistics at Birkbeck ended in 1962 when Booth left England for the University of Saskatchewan. There, some research on MT was taken up by his wife, Kathleen Booth, on English-French translation (ch.12.5), and this was continued in some form when the Booths moved to Lakehead University, Thunder Bay, Canada.

5.2. Cambridge Language Research Unit (1956-1967)

Research on MT at Cambridge began, as we have seen (ch.2.6), with the informal meetings of the Cambridge Language Research Group in 1954. Considerable interest was aroused by the originality of the approaches put forward at the August 1955 meeting in King's College, Cambridge. In 1956 a grant was received from the National Science Foundation to pursue MT research, and the Cambridge Language Research Unit (CLRU) was formed, a research organization independent of the University of Cambridge, with Margaret Masterman (Mrs. Braithwaite) as director.[2] In later years research grants were also received from the U.S.Air Force Office of Scientific Research, the Office of Scientific and Technical Information (London), the Canadian National Research Council, and the Office of Naval Research (Washington, D.C.) Although MT was the primary interest in the earlier years, other fields of research were and have been pursued vigorously at the same time: information retrieval, automatic classification, computer simulation, on-line computer interaction, and recently 'breath-group' analysis. (The bibliography in Masterman (1970) reflects the great variety in CLRU research interests.) By 1967, active research on MT at CLRU had declined; many of its members had moved elsewhere, mainly within the Artificial Intelligence field and often with a continued interest in MT (e.g. Martin Kay and Yorick Wilks, ch.17.8 and 15.1 below). However, the study of MT problems has continued at CLRU to this day, albeit at a much lower intensity than during the 1950's and 1960's.

The Cambridge (CLRU) group has been characterised throughout by a diversity and prolixity of theories and methods.[3] There were four main themes in its MT research: the thesaurus approach, the concept of an interlingua, 'pidgin' translation, and lattice theory. The focus was primarily on semantic problems of MT, and syntactic questions were treated secondarily. Although procedures were intended to be suitable for computers, most of the proposals were tested only by manual or punched card simulations because access to a computer proved to be difficult for many years.

The principal objective of CLRU was the investigation of methods which would, in the long term, produce good-quality fully automatic idiomatic translations. The fundamental problem was recognised to be that of polysemy or 'multiple meaning'. The ultimate solution was believed to be translation via an interlingua, but it was recognised that considerable semantic research had to be undertaken before even the outlines of a genuine interlingua could be discerned. The CLRU adopted two basic lines of research: the development of a crude prototype interlingual system producing 'pidgin' (essentially word-for-word) translations, and the development of a complex, sophisticated tool for improving and refining the lexical expression of unsatisfactory MT output, dealing particularly with problems of polysemy. In both lines of research, a central role was played by the notion of a thesaurus (a structured conceptual (semantic) classification of vocabulary), as the tool for output refinement and as the basis for an interlingua.

Thesauri, of which *Roget's Thesaurus* is the most familiar example, classify vocabulary into groups of words (and phrases) having similar meanings and arrange them under a number of 'heads'. These headings may be interpreted as the contexts in which the listed words may occur;

[2] For an assessment of her MT research see Y.Wilks: 'Margaret Masterman', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 279-297.

[3] See also J.Léon: 'Traduction automatique et formalisation du langage: les tentatives du Cambridge Language Research Unit (1955-1960)', *The history of linguistic and grammatical praxis*: proceedings of the XIth International Colloquium of the Studienkreis "Geschichte der Sparchwissenschaft", (Leuven, 2nd-4th July 1998), ed. P.Desmet et al. *Orbis Supplementa* 14 (Leuven: Peeters, 2000), 369-394.

and in principle, such contexts are extra-linguistic, i.e. not language specific but interlingual. Any particular word may (and often does) appear under more than one 'head', i.e. in more than one context, according to its different senses. Thus, *work* would appear under both Intellectual labour and Manual labour. The particular context would determine which 'head' was relevant: in *He's been working on that problem for two months* the occurrence of *problem* would indicate Intellectual labour. In abstract (mathematical) terms, the structure of thesaural interrelationships could be defined as a lattice, in which every word could be located according to its presence or absence under each of the thesaural 'heads'. The potentialities of applying lattice theory to linguistic patterns were examined in great depth by the CLRU team, not only for semantic (thesaural) operations but also for syntactic operations (e.g. Parker-Rhodes 1956, 1956a, 1961).

In one line of CLRU research, the thesaurus approach was investigated as a method of improving the unsatisfactory translations produced by existing MT systems. These word-for-word translations were to be treated as interim 'pidgin' English versions, i.e. following Richens' view of his and Booth's early efforts (ch.2.2 above) and Reifler's comments on the Washington University MT output (ch.4.1 above). Such 'pidgin' versions were to be made idiomatic by the additional operation of a thesaural 'retranslation'. The thesaurus would provide access to the rich variety of synonyms and idiomatic usages which could not be incorporated in bilingual dictionaries. In the thesaural approach then, the choice of appropriate TL version for a particular 'pidgin' word was envisaged as first a search for the correct 'head' (context) and then selection from the list of synonyms (Masterman 1957). For example, the word *plant* would be found to have a number of different contexts, e.g. plant as place, 184: as insert, 300: as vegetable, 367: as agriculture, 371: as trick, 545: as tools, 683; etc., each number standing for a list of synonyms which might appear in the context of *plant*. Initially, it would not be known which of these lists of synonyms of *plant* should be chosen. However, if *plant* was preceded in the text by *flowering* then consultation of this word in the thesaurus would give another set of synonym lists, e.g. flower as essence, 5: as produce, 161: as vegetable, 367: as prosper, 734: as beauty, 645; etc. Only one of these lists is common to both *flower* and *plant*, namely the list under the 'head' Vegetable; and this is clearly the correct context. In many cases there would be more than one, and selection would then involve searches for words common to all the synonym lists. This procedure was applied to the translation of Italian *alcune essenze forestali e fruttiferi*, viz. FOREST AND FRUIT-BEARING ESSENCE-S. The thesaurus heads for *forest, fruit, bearing* and *essence* were consulted: *fruit* and *forest* both appear under Vegetable, *fruit* and *bearing* both under Production, *bearing* and *essence* both under Meaning, etc. A complex process of comparison of lists established links between Production and Meaning via Intrinsicality and Prototype, and produced *example, specimen, pattern, prototype* as alternatives of ESSENCE (Masterman 1956).

The parallel line of CLRU research continued Richens' ideas on interlingual translation (Richens & Booth 1955, Richens 1956a, 1956b, 1956c).[4] It started from the basic distinction between lexical items (stems) and grammatical 'operators' (e.g. endings or function words). Lexical items were to be transferred via a crude interlingual dictionary of 'naked ideas' (Nude), semantic elements structured as a thesaurus. The operators were also to be analysed into interlingual functional categories, e.g. 'used to indicate past time', 'used to indicate inanimate objects'. (Some of this research was the work of Halliday (e.g. Halliday 1956) and may perhaps be seen as containing seeds of his functional approach to grammatical description in 'systemic grammar', e.g. Halliday 1973, 1985)

The result of the operator analysis was a 'syntactical thesaurus' structure, i.e. groupings of operators under 'heads' such as In/animacy. The fact that such heads might also be used for lists of lexical items, and the fact that the analytical method proceeded by dichotomous cuts, strengthened

[4] For Richens and the CLRU interlingua see K.Sparck Jones: 'R.H.Richens: translation in the NUDE', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 263-278.

the conviction that language systems were basically lattices. However, the 'thesaural' approach to syntax was found less satisfactory than the alternative of a simple classification of grammatical functions as noun-like or verb-like or qualifying functions, together with a simple 'bracketting' procedure for indicating structural relations (a 'syntactic lattice').

The procedure for 'bracketting' was described by Parker-Rhodes (1961, 1962, 1966). Classification of word classes, phrases and clauses was based on dependency relations, and individual words were coded according to their potential occurrences as governors, dependents, either or neither in each type of group (coordinate structure, adverbial group, participial clause, nominal group, etc.) Analysis consisted in the examination of sequences in order to identify potential governors and potential groupings ('brackettings'). An example analysis was:

((a (rather lazy) cat) (chases (falling (leaves and butterflies;)))) (of course these
(can (easily get away.)))

Bracketting was the major method of defining boundary conditions for thesaural operations.

Semantic analysis, however, was throughout the focus of CLRU research. Much work was done on improving thesaural structures. A major problem was the very notion of synonymy. Sparck Jones (1962, 1965) investigated in depth the question of devising automatic methods of semantic classification. This research was closely linked with the extensive studies at CLRU of semantic aspects of information retrieval, where again the thesaurus concept was central (Needham & Joyce 1958), and which suggested fruitful analogies with MT systems (Masterman et al. 1959).

Probably the most important contribution of the CLRU team was to explore more thoroughly than ever before the theoretical and practical difficulties of constructing an interlingua based on a finite set of primitive semantic components. Much research effort was expended on the building of Nude (interlingual) dictionaries, primarily for translating Italian and Latin. The main difficulty was the establishment of the interlingual semantic components themselves. Evidently, thesaural 'heads' alone were insufficient. More refined analysis was necessary in order to distinguish between rows of synonyms. One possibility explored was the inclusion of primitive classifiers (derived ultimately from iconic and 'logical' categories.) The classifiers proposed (Masterman 1961) were suggested in part by Ivor Richards' *Language Through Pictures* series for teaching languages (e.g. Richards & Gibson 1952-58). They included HE, SHE, I, YOU, DO, BE, BANG (suddenness), ONE, PAIR, KIND, HOW, CAUSE, CHANGE, MORE, LESS, FOR, SPREAD, MAN, THING, FOLK, BEAST, PLANT, etc. Such classifiers, in suitable combinations, could subdivide thesaural lists; e.g. the 'head' 839 LAMENTATION:

KIND	lamentation, mourning, grief, sobbing, tears
ONE BE	sob, sigh, complaint, whine
BANG KIND	flood of tears, crying, howling
ONE BANG BE	outburst of grief, cry, scream
THING	weeds, crepe, passing-bell, dirge, wake, funeral
SHE THING	widow's weeds
MAN DO	mourner, weeper
DO	lament, mourn, fret, groan
MORE DO	burst into tears
LESS DO	sigh, shed a tear

and so forth (Masterman 1961)

Augmentation of thesaural lists in this way could serve a further purpose by adding message-structuring information ('syntactic' information in a broad sense). For example (Masterman 1962), the semantic gist of the sentence:

This man can eat, all right; but he can't, for the life of him, fight

might be expressed by the sequence of 'minimal semantic units', plus indications of basic operators (colons for noun-like functions, slashes for verb-like functions) and bracketting:

(THIS: MAN:) ((HE: (CAN/ DO/ (MUCH: EAT:))) (BUT: NOT:) (HE: (CAN/ DO/ (MUCH: FIGHT:))

The closeness in meaning of this sentence to:

This man is greedy, but pusillanimous

would be shown by the similarity of the 'semantic message' analysis of the latter (omitting syntax):

THIS: MAN: HE: MUCH: WANT/ EAT/ BUT: HE: SMALL: WANT/ FIGHT/

The complexity of the thesaurus and the analysis procedures for producing such 'semantic shells' was recognised to be well beyond the capabilities of any then (and perhaps even now) conceivable computer. In addition, of course, a MT system would need complex bilingual dictionaries (from SL to interlingua, and from interlingua to TL.) The semantic universe is so enormous that in "compiling a realistic dictionary for MT, the scale of a bi-lingual large Oxford English Dictionary (18 volumes) would be far too small; a 200-volume dictionary would be more like what would be required" (Masterman 1962) Evidently, while the long-term aim might be the development of a thesaurus-based interlingua, there had to be more limited practical research objectives.

The further investigation of 'pidgin' translations was one obvious option. The idea was to test "the whole Mechanical Pidgin idea to destruction, in order to see what can be done with it and what cannot" (Masterman & Kay 1960). The characteristics of 'pidgin' were held to be: the predominant use of phrases, rather than words, as dictionary units; the employment of specialised dictionaries; the use of constructed symbols, e.g. 'pidgin variables' such as (W)THAT and 'pidgin' grammatical markers such as -ISH and -WARD for adjectives; the avoidance of certain problems of translation (such as articles, case endings and prepositions); the provision of ideally just single equivalents (no alternatives indicated); and strict adherence to the word sequence of the original. The thesis propounded was that 'pidgin' translation represented absolute minimal MT, and that consequently no more elaborate system could be justified unless it produced translations "noticeably better than Mechanical Pidgin" (Masterman & Kay 1960)

Experiments were made with Latin-English, producing the following for Julius Caesar's familiar *Gallia est omnis divisa in partes tres...*:

> Gaul is all divided in part(s) three, of which one inhabit-they the+Belgians, the+other the+Aquitains, third one of+themselves language the+Celts our the+Gauls call-they+are/-they-would. These all language, custom, law-s from+one+another differ-they. The+Gauls from the+Aquitains the+Naronne the+river, from the+Belgians the+Marne and the+Seine divide-s...

As a second experiment, the 'pidgin-improving' devices developed were applied to a (unedited) Russian-English translation of a *Pravda* article produced by the IBM system (ch.4.2); Gilbert King was at this time a consultant of the CLRU project and a believer in the validity of the CLRU approach (King 1958). First, an extract from the IBM version:

> In this meaning very urgent located in magazine article about first *entsiklike* (message) chapter Roman-Catholic church *papy Ioanna* XXIII with/from that it inverted at the end June present year to *episkopan*, sacred*kan* and believing, consisting in Catholic church.

The CLRU 'pidginized' version was:

> *From+this+point+of+view* very timely-is which+is+located in journal - - article - - about first encyclical-? (message-?) head-'s Roman-Catholic church-? pope-'s John XXIII with/from which that+one address-ed *at+the+end* June-ish present year-'s to bishop-s-ward priest-s-ward and to+believers, *belonging+to* Catholic church-?

The justification for investigating 'pidgin' MT was twofold. Firstly, most MT output was (and was likely always to be) non-idiomatic and in need of some semantic refinement, i.e. via a CLRU-type interlingual thesaurus. Secondly, investigation of 'pidgin' versions compelled close examination of the ‘bits of information’ contained in the texts, the semantic phrasing or patterning of the message; it forced MT research to concentrate on its primary goal: the transfer of the meaning of messages (Masterman 1967).

Further insight on semantic phrasing was obtained from work by Shillan (1967) on the articulatory segmentation of texts, the natural phrasing of messages into information blocks containing on average two stress-points. For example:

> Canada is divided/ into ten provinces/ and two territories./ Each province is sovereign/ in its own sphere/ and administers its own/ natural resources./

The segmentation of texts in this way suggested that phrase-by-phrase translation might be a more ‘natural’ approach than traditional sentence-by-sentence translation. However, in addition, further light on basic message structures was suggested by the notion of ‘semantic squares’ (Masterman 1965), a theory describing a patterning of semantic contrasts, reiterations and recapitulations which would appear to underlie argument and paragraph structures, and which transcends specific language structures. The basic unit of a 'semantic square' was a coupling of phrases with two stress points. The relationship between stress patterns (‘breathgroups’), reiteration and semantic information structure has remained a focal point of Masterman’s research to the present day (e.g. Masterman 1982)

In the mid-1960’ the CLRU group investigated the possibilities of interactive English-French MT for the Canadian National Research Council (Masterman 1967a). The proposal was a machine aided system in which the translator is on-line to a thesaural dictionary from which lexical items are retrieved by ‘requests’ formulated in ‘natural phrasing’ units and with which the translator interacts by question and answer routines. For example, to translate HE FLEW TO THE FRONTIER the operator began by entering the phrasing-frame:

```
operator: HE xx-ED TO THE xx
machine: DO YOU MEAN
        A HE MOVED TO SOME POINT?
        B HE CROSSED SOME OBSTACLE?
        C HE COMMUNICATED WITH SOME BEING?
operator: A
machine: DO YOU MEAN
        A HE DEPARTED TO SOME DESTINATION?
        B HE TRAVELLED TO SOME DISTANT REGION?
        C HE FLED TO SOME REFUGE? etc.
operator: B
machine: HOW DID HE TRAVEL?
        A: BY PLANE?
        B: BY SEA?
        C: BY ROAD? Etc.
operator: A
```

The operator then typed in the two stressed words: FLEW and FRONTIER and the machine would print out:

```
IL PRIT L’AVION POUR LA FRONTIÈRE
```

Included among the techniques was a method derived from the notion of ‘semantic square’. The translation of HE WENT TO THE POLICE established first the option HE COMMUNICATED WITH SOME ANIMATE BEING, and then HE REVEALED-ALL TO THE ENEMY (rather then ‘told a story’ or ‘consulted with some authority’). A matching of POLICE and ENEMY produced the French *les flics* (whereas the matching of POLICE and AUTHORITY would have produced *la police*).

The project was important as one of the first to investigate interactive MT (and in comparison with later systems it was certainly crude) but it was believed to have added significance in that on-line MT could reveal the actual methods used by human translators. Insight into actual translation processes ought to lead to improved automatic systems and aids of genuine value to practising translators (cf. 19.5, and Masterman 1979).

The CLRU research team saw its approach (primarily the emphasis on interlingual thesaurus research) not as a rival to other MT systems, but as complementary. It concentrated on those extremely difficult and complex areas of semantic which other projects tended to shy away from. For this it was criticised as being too idealistic. (A particularly sharp attack on the CLRU approach by Bar-Hillel (ch.8.3 below) was answered by members of the group in a compendium of CLRU research: *Essay on and in machine translation* (CLRU 1959).) The fact that no MT system as such emerged from CLRU research is an irrelevance. In recent years, research in Artificial Intelligence has turned increasingly to the areas of investigation which were first examined in depth by the Cambridge project. Some of CLRU research was found to be on unfruitful lines; perhaps the ideal of a genuine interlingua was shown conclusively to be a chimera. On the other hand, features of the CLRU conception of semantic message structure have lived on in various guises in both AI and MT research (cf. Ch.15 below).

5. 3: University of Milan (1959-1966)

The approach of the MT research group in the Centro di Cibernetica e di Attività Linguistiche, Milan, headed by Silvio Ceccato, had the long-term objective of establishing a MT method based on the supposed thought processes of human translation.[5] Translation was conceived as the passage from SL to TL via mental constructs underlying the text in question. The goal of linguistic analysis was to study the relations between thought processes and expression, and to establish the basic structures of thought content. “A man who is translating is thinking; his understanding of the original text is thinking, and his translated text designates his thoughts.” (Ceccato 1966). In brief, it was an 'interlingual' approach to MT, but one which, Ceccato was always at pains to stress, did not start from linguistic principles but from philosophical foundations. Indeed, Ceccato prefaced all his substantial articles by detailed expositions of his philosophical premisses (Ceccato 1961, Ceccato & Zonta 1962, Ceccato 1966, Ceccato 1967)

Ceccato was a member of the Italian Operational School of philosophy (Scuola Operativa Italiana, established in 1939), which had been engaged for a number of years in developing models of mental activities, one mechanical model Adamo II being exhibited at an Automatism Exhibition in 1956 in Milan (Albani et al. 1961) In 1955 he and Enrico Maretti, an engineer responsible for Adamo II, were invited to present their nascent ideas on MT at the ‘Information Theory’ conference in London (Ceccato & Maretti 1956). Ceccato continued to develop his ideas until in February 1959, a grant was received from the US Air Force Research and Development Command (Rome Air Development Center), through its European Office of Aerospace Research, to engage in MT research primarily towards Russian-English translation (US House of Representatives 1960). At a later date the project was also reporting to EURATOM (Ceccato & Zonta 1962). The group conducted almost exclusively theoretical studies, with occasional small-scale simulations, examining Italian, German and Latin as well as English and Russian, and remained active until about 1966 (Josselson 1970).

The primary argument of the Italian Operational School of Philosophy was that the contents of thought should be regarded as activities and not, as in traditional philosophy, as objects (whether observational objects or characteristics of observed objects). Instead of analysing thought in relation to already extant ‘things’ or ‘concepts’, they proposed to study thought and its contents in terms of the operations required to mentally construct these contents. In this way, they claimed to eliminate the mind-body and concrete-abstract dichotomies. Four fundamental operations were

[5] See also E.von Glasersfeld: ‘Silvio Ceccato and the correlational grammar’, *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 313-324.

identified: differentiation, figuration, categorization, and correlation. Differentiation was defined as the activity by which changes of state are perceived and which “allows us to speak, for example, of warm and cold, of light and darkness, of hard and soft, of attention and inattention, of silence and noise, of good and bad moods, etc.” (Albani et al. 1961). Figuration was defined as the activity of constructing forms and spatial localisation; in conjunction with differentiation, it gives rise to perception and representation. Categorization was defined as the activity of mental classification, which “gives us the mental, or logical, categories, including, for example, substance, accident, subject, object, and, or, with, also, by, state, point, line, surface...” (Albani et al. 1961); it was regarded as an operation based on memory of temporally distinct differentiations. Correlation was the activity of thought itself, relating or ordering “material received from... other activities”. It was defined as a triadic relation of a correlator (a mental category) and two correlata (either other mental categories or the results of other types of activities). Representations of the thought content of sentences or phrases were given as networks of correlational triads.

It was Ceccato's contention that traditional linguistics was inadequate for MT; it could not deal with discontinuous structures (e.g. that in: *In a deck chair, worn and depressed, there sat a young woman* we know *worn and depressed* must refer to *woman* and not to *deck chair*), or with homography and polysemy (e.g. “the four different situations” designated by *with* in *to sing with grace, to sing with Callas, to sing with a microphone, to sing with the dawn*). This was because grammar “was born not so much as a key for interpreting discourse, and thus to pass from the words to the designated thought; but rather to systematize language... In any case, grammar presupposes that whoever uses it already knows how to think...” (Ceccato 1967) What was needed, therefore, was “to build up a new grammar, with research going far beyond the normal classifications of words and the normal rules which suffice for the guidance of humans” (Ceccato 1966) In particular, it was considered important that ‘intermediary’ representations of thought content were not to be determined by structures of particular languages or of quasi-linguistic (including logical) forms (Ceccato 1961), because “Languages are not fabricated from one another, but are all built up on thought and must contain the indications to express it” (Albani et al. 1961). Ceccato was consequently opposed to the idea of an interlingua based on ‘universal’ or ‘common’ linguistic features.

Ceccato's method had the following stages. First, “correlational ‘tabellone’” are set up: “For each word, the correlational possibilities of the thing which it designates are examined”, e.g. “‘water’ designates a correlatum, either first or second, of a correlation whose correlator is ‘and’, ‘or’, ‘with’, etc.; but it cannot be the second correlatum of a correlation whose correlator is ‘between’, because a plural is required for this position; we can say ‘water and’, ‘and water’, ‘water or’, ‘or water’, ‘water with’, ‘with water’, ‘water between’ but not ‘between water’”. The set of correlators themselves were considered to be “relatively few in number even in the richest and most highly-developed trains of thought: we may say between 100 and 200 in all”. They included the conjunctions and prepositions, punctuation marks, and relations such as subject-predicate, substance-accident (i.e. noun-adjective), apposition, development-modality (i.e. verb-adverb), and comparison (Albani et al. 1961). The intention was that the correlators should be universals, but as a “concession to practical considerations” the table of correlators was “designed instrumentally, as a linguistic-correlational chart”, one for each of the languages Russian, English, Italian, German and Latin.

After every word has been classified by its correlational possibilities (quite a task in view of the fact recognised by Ceccato (1966) that “for certain words, nearly all the possibilities are open. The word ‘water’, for instance, will occupy more than 160 positions.”), the next stage was to establish the conditions which apply, or had to be satisfied, for each possible correlation. Some of these limitations were general (e.g. that if one correlatum had been identified, the other must be found); others were specific (e.g. adjective-noun and subject-verb agreements). The final stage of grammar construction was the classification of the correlations themselves according to their potentialities as correlata within other correlations.

To deal with polysemy, Ceccato proposed "a much subtler analysis than the one carried out so far in relation to the tabellone; we now need an analysis of what the words designate" (Ceccato 1966). For this he suggested classifications of, for instance, *apple* as "an observatum, physical, with a form, and... a specimen of the class of edible things", and the construction of 'notional spheres' in which 'developments' (i.e. verbs), 'things' (nouns) and their 'aspects' would be linked by 'basic relations'. It was claimed that the set of these relations was limited, "not more than a few hundred" and Ceccato (1966) listed 56 such as 'member: part', 'species: genus', 'part: whole', 'thing produced: thing which produces it', 'thing produced: place of production', 'thing contained: container', 'thing supported: support', etc. Thus *sleep* was linked to *bed* (activity: usual place), and to *night* (activity: usual time), and *bed* was linked to *bedroom* (object: usual place), and to *furniture* (species: genus), and so forth.

The understanding of a sentence (or text) involved, therefore, the construction of a correlational net on the basis of information about the correlational possibilities of each word, the possible linkages among correlations, and limitations on relations within the 'notional sphere'. Translation involved the transformation of a correlational net into an equivalent net appropriate to the correlational possibilities of words and correlators of the target language. Such transformations would include, for example, the insertion of articles and the completion of elliptical constructions (e.g. for Russian-English translation, cf. Perschke in: von Glasersfeld et al. 1962).

Although perhaps a number of simulations were performed – Zonta states: "input vocabulary for our current programme includes about 50,000 inflected Russian forms, corresponding to about 2,500 headwords and punctuation marks" (Ceccato & Zonta 1962) – only three "microexamples" of translations are reported, of the Italian sentence *Un giglio ci sta bene* into English (*A lily goes well*) (Ceccato 1966), and of the English sentences *Were I tired, I should weep* and *Engineer Small has a little train* into Italian (Ceccato 1961: 175ff.; Ceccato 1967).

These examples confirm the conclusion evident from this outline that correlational analysis was no more than a version of immediate constituent analysis (ch.3.4), based on traditional grammatical categories and relations (nouns, prepositions, subject-predicate, verb-adverb), and constrained by limits on the collocation of semantic features (also of a traditional nature: 'physical', 'edible', etc.) For example, the analysis of *Engineer Small has a little train* gives the correlational net (Ceccato 1967):

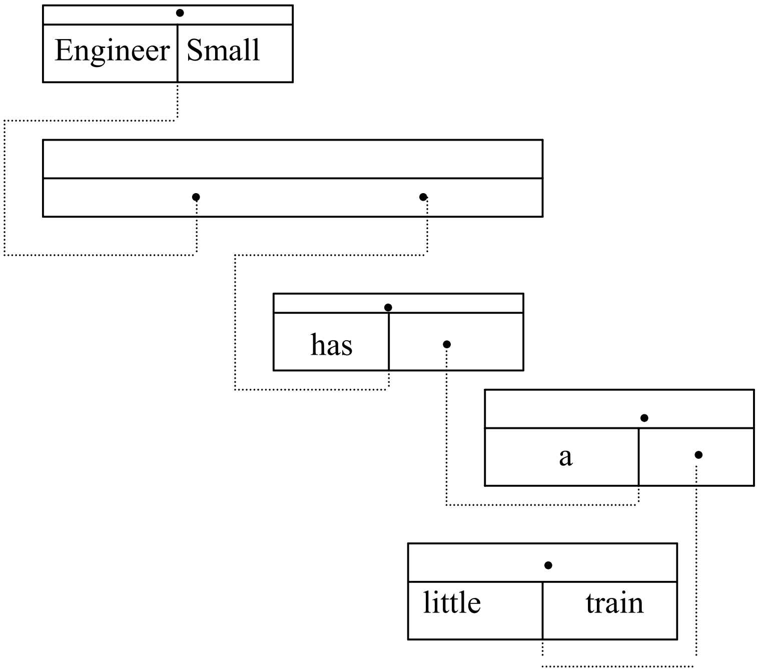
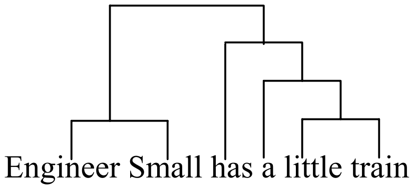


Fig.11: Correlational net

Which may be seen as equivalent to the immediate constituency analysis:



Only the idea of ‘notional spheres’ represented a new departure, although even in these the types of relations were familiar, and their application in analysis was not described. In theory, analysis was to take into account intersentential relations, but in practice analysis was restricted to sentences as units. Finally, the transformations of correlational nets (roughly, phrase structures) in translation were not much different than other syntactic transfer routines of the time.

From the viewpoint of practical achievement in linguistic analysis, Ceccato's method was no advance on traditional approaches. He dismissed contemporary linguistics, but in effect his 'philosophical' approach reinvented traditional grammatical categories and classifications. His introspection had discovered not the universal processes of thought but the basic categories of the language in which he thought (Mounin 1962); as indeed he virtually admitted, when describing the modifications to correlational analysis prompted by German *damit, davon, danach*, etc. (Ceccato 1966) As a result, correlational grammar was effectively just another version of phrase structure grammar, as Glasersfeld & Pisani (1970) confirmed when at Georgia University they programmed a small-scale experimental parser, the ‘Multistore’ parser, based on the correlational approach.

What is more important, in retrospect, was the cognitive orientation to MT and text analysis in general. Ceccato emphasised the dynamic mental operations involved in verbalisation and comprehension. Ultimately the goal was a machine “capable of carrying out some of the human operations of observation, mental categorisation, thought and language” (Albani et al. 1961). Genuine MT, Ceccato believed, required the construction of a 'linguistic machine', "a machine which would follow a discourse, carrying out all the operations which, in us, constitute understanding” (Ceccato 1966); and would include “the construction of a memory with the characteristics of human memory, which is associative, selective, and propulsive" (Ceccato 1967). In other words, it would appear that his primary interest was not MT (although this was seen as a practical and feasible goal) but what is now called Artificial Intelligence; Ceccato was convinced that despite the enormous difficulties, it was "desirable to continue our study of thought and language in man; both for the general theoretical value of these studies and for their contribution, considerable in the past, and certain to increase in the future, to the project of the mechanization of intelligent activity" (Ceccato 1966).

Limited and tentative though they were, we may regard aspects of Ceccato's cognitive (and interlingual) approach as genuine foreshadowings of some future AI methods (ch.15 below); his 'notional spheres' have obvious analogies with semantic networks; correlational nets are similar (in conception if not in form) to the conceptual dependency networks of Schank; and the correlational analysis of sentences and texts may be seen (again in conception if not fully in practice) as an early form of semantics-based parsing.

5. 4: National Physical Laboratory, Teddington (1959-1966)

The Autonomics Division of the National Physical Laboratory (NPL) at Teddington, near London, began work on a 'pilot' Russian-English MT system in 1959 (McDaniel et al.1967, 1967a, 1967b; Szanser 1966, 1967). The research formed part of a wider programme of investigations into automation and robotics. From the start, the aims were both practical and limited: a demonstration of the practicality and feasibility of Russian-English MT of scientific and technical texts for the expert reader. The research ended in 1966, when the computer used throughout the project was scrapped. This was the experimental machine ACE designed and constructed at NPL as a development of Turing's pioneer work at the NPL just after World War II.

The basic dictionary was obtained initially from a copy of the Harvard University Russian-English dictionary. This was revised and adapted, and ultimately contained about 15,000 words represented by 18,000 entries. As in the Harvard dictionary (ch.4.9 above), stems and endings were entered separately except in the case of irregular forms which were left unsplit. Some differences in the representation of Russian morphological forms were introduced (McDaniel & Whelan 1962), with consequential extra work which meant the dictionary was not fully operational until 1963 (McDaniel et al.1967a). As in most systems at the time, dictionary lookup was serial, matches being made against stems on the longest match principle and idioms being identified and translated as wholes. If a word did not appear in the dictionary, some attempt was made to 'anglicise' the Russian by two routines, one which transliterated stems (on the argument that many new scientific words in Russian have 'international' stems), and another which identified common international and Russian prefixes and supplied English equivalents, e.g. *radio-*, *elektro-*, *mnogo-* (*many*), *poly-* (*semi-).*

It was intended to introduce progressive refinements of the basically word-for-word translation by syntactic operations. A number of procedures were investigated, but few were implemented before the end of the project (Szanser 1966; McDaniel et al. 1967a). The first refinement was the recognition of 'nominal blocks' (noun phrases with adjectival and noun modifiers) in order to resolve homographs via the establishment of word classes, and to identify places for the insertion of English prepositions. A second refinement was a similar procedure for predicate blocks (finite verbs and modifiers), including recognition of reflexive verbs in order to ensure output in appropriate English passive forms. Further syntactic routines were simulated satisfactorily but not implemented. These included the identification of clauses by location of conjunctions, relative pronouns, etc., the identification of coordinate structures, and the recognition of subject-verb and verb-object government. Other syntactic procedures investigated included the resolution of ambiguities, such as case ending ambiguities (including insertion of English prepositions), adverb and short adjective ambiguities (e.g. whether Russian *tochno* is to be translated *is accurate* (short adjective) or *accurately* (adverb)), and third person pronoun ambiguities (e.g. Russian *ego* as *his*, its, *him* or *it*); and finally some study was made of the treatment of ellipsis (e.g. the common omission of the copula in Russian).

In view of the low level of syntactic analysis actually incorporated, the system produced translations which were little better than word for word versions. An example from Szanser (1966):

> THESE MUTUAL INTERFERENCES HAVE OWN REASONS AND/ALSO
> DEPEND ON POWER/OUTPUT, QUANTITY(S) AND/ALSO RANGE(S)
> OF TRANSMITTERS FROM RECEIVING EQUIPMENT,
> LOCATION/ARRANGEMENT OF THEIR AERIALS, DIFFERENCE(S) OF
> FREQUENCIES OF TRANSMITTERS OR THEIR HARMONICS FROM
> FREQUENCIES OF RECEIVERS AND/ALSO, AT LAST/FINALLY, FROM
> INTENSITY OF RADIATION TRANSFERRING AND/ALSO
> AMPLIFICATION/MAGNIFICATION/GAIN OF RECEIVING DIRECTED
> AERIALS IN NOT WISHED DIRECTIONS/TRENDS.

The example illustrates also the lack of any semantic analysis (beyond the limitation of vocabulary to specialised fields), although again some plans were being made at NPL before termination of the project (Szanser 1966).

Intended always as an experiment in MT feasibility, it was fitting that the project concluded just as the operational system had reached a stage when its practical usefulness could be evaluated (Szanser 1967). In May 1966, practising scientists in British universities, government research institutes and industry were invited to submit Russian articles for translation. In general, the results were received favourably; a grading of users' comments in terms of perceived 'usefulness' produced a mean rating interpretable as "slightly less than 'good'" (Szanser 1967). However, the inadequacies of the lexicon and the inclusion of too many alternatives were points of criticism. (McDaniel et al. 1967b).

Although the theoretical contributions of the NPL team were negligible and the operational system represented little advance on basic word-for-word systems, the NPL project did produce translations which could be evaluated by scientists in a practical environment. It remains to the present day the only operational MT system developed so far in Great Britain.

5.5: Research in France (1959-66)

In December 1959 the Centre National de la Recherche Scientifique (CNRS) established a MT research body, the Centre d'Etudes de la Traduction Automatique (CETA), at two centres in Paris and Grenoble, under the general direction of Peres (*CRDSD* 8, May 1961; *TA* 1(1), April 1960).[6] Both centres were to work towards a system for translation from Russian into French. Within one year the centres employed 25 research workers (*TA* 2(4), 1961) According to the US House of Representatives report (1960) contact with US research was maintained by a "member of the U.S.Air Force" who served "as the U.S. representative on the mutual weapons development team, for exchange of technical data on automatic language translation research within France".

At Paris, the team under A.L.Sestier was engaged on the study of the syntax of German and Russian; a successful trial syntactic analysis of Russian on an IBM 650 was reported in September 1961. Most of its activity was devoted to problems of dictionary compilation and searching (e.g. Meile 1962), and by 1962 the Russian dictionary contained 12,000 stems and multiple paradigmatic classes had been established (Dupuis 1966) However, not long afterwards, the Paris section ceased involvement in MT research and from 1963 onwards all CETA activity was centred at the University of Grenoble.

[6] See also J.Léon: 'Les débuts de la traduction automatique en France (1959-1968): à contretemps?', *Modèles Linguistiques* 19(2), 1998, 55-86; and M.Gross: 'Early MT in France', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 325-330.

The Grenoble group is now one of the longest established MT groups, directed by Bernard Vauquois since its creation in 1960 (Vauquois 1966a).[7] An early decision of this group was that MT could be successful only if analysis was pursued to a much 'deeper' level than most contemporary systems. It was acknowledged that at the current stage of MT research semantic analysis was not a practical proposition and that consequently most attention had to be paid to ‘deep syntactic’ analysis, and to methods of syntactic transformation. The aim was to produce representations which would serve as the source for TL synthesis; in other words, CETA was working towards an interlingual representation (at least as far as syntax was concerned), a conception which was later to be termed a ‘pivot language’. A particular long-term emphasis of the CETA group at Grenoble was the great attention paid to the establishment of powerful algorithmic programs based on rigorous modelling of linguistic formalisations. The aim became to “établir le système de traduction automatique au moyen d’une succession de modèles logico-linguistiques” (Vauquois 1966a). This emphasis lead to substantial investigations of algebraic and formal linguistics (e.g. Vauquois’ 1962 paper (Vauquois 1966) concerned primarily with problems of morphology), and to the development of the notion of ‘sub-grammars’ to increase the algorithmic efficiency of analysis programs (Vauquois et al. 1965).

The Grenoble group began research on three MT systems for Russian-French, German-French and Japanese-French. The latter was brought to an early end by the departure of Yamada in July 1962 (cf.ch.7.1 below), but it was resurrected at intervals during later years. The Russian-French system received highest priority; by 1966 the linguistic foundations for morphological and syntactic analysis were said to have been completed (Vauquois 1966). The German-French system was also said to be progressing, though with much less urgency than the Russian-French project. For their Russian dictionary, the Grenoble group was able to build on the corpus of some 700,000 words compiled by RAND (ch.4.4)

Although it is evident that some of the principal characteristics of this MT group had already been formulated by 1966 (notably the syntactic interlingua and rigorous formalism), research was still at an early stage and had not (except at a theoretical level) had much impact on the general direction of MT research at this period. As we shall see (ch.10.1 and 13.3), the Grenoble group was to be most influential after research in the US had been interrupted by the ALPAC report (ch.8.9 below).

At the same time as the Paris and Grenoble groups were set up, a MT group began in 1962 at Nancy composed of Bernard Pottier, Guy Bourquin and Legras, also sponsored by the CNRS. There was some talk at the time that this group should also form part of the CETA project, to be known as the ‘Section de Nancy’. However, it remained independent. A successful experiment on English-French translation was reported in late 1960 (*TA* 2(4), Dec 1961); however, this must have been a very tentative exploration as it is evident that MT research at Nancy has always been of a long-term theoretical nature: statistical analyses of English scientific vocabulary (including semantic studies), some studies of French syntax (particularly by Pottier), and similar work on Spanish (*TA* 5(1), 1964). Although always including MT within its purview, the Nancy group has had wide interests in many areas of computational linguistics. It remains active to the present day (cf.ch.14.2 below).

There was also a short-lived group (1961-62) set up by the Association pour l'étude et le développement de la traduction automatique et de la linguistique appliqué (ATALA), and led by Michel Corbé of Unesco and Robert Tabory of IBM-France. The aim was an English-French system, basically of the ‘syntactic transfer’ type and influenced by Yngve’s work at MIT (ch.4.7). The methodology was based on the 'empirical' approach at RAND and Ramo-Wooldridge (ch.4.4

[7] For Vauquois and research at Grenoble see: *Bernard Vauquois et la TAO, vingt-cinq ans de traduction automatique: analectes*, ed. C.Boitet, Grenoble: Association Champollion & GETA, 1988; and also C.Boitet: ‘Bernard Vauquois’ contribution to the theory and practice of building MT systems: a historical perspective’, *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 331-348.

and 4.6), and there were hopes of using a photoscopic memory device of the IBM type (ch.4.2). The preliminary work on English syntax was reported by Corbé & Tabory (1962). The proposed parser would identify sentence fragments and determine their interrelationships in a series of four passes. There is no evidence that the system was implemented.

5.6: Research in Belgium (1961-64)

At the University of Brussels a project was set up during 1961 under the leadership of Lydia Hirschberg (*TA* 2(4), Dec. 1961). Funded by EURATOM its aim was the investigation of methods and procedures for Russian-French MT. The approach adopted had similarities with Harris' string analysis and with dependency grammar (Josselson 1970). A good deal of theoretical work was accomplished, but it seems that only morphological procedures were actually implemented (Blois et al. 1968). From 1962 the group worked closely with Euratom's computer centre CETIS on the adaptation of the Georgetown system for Russian-French MT, leaving the GAT Russian analysis program unmodified and developing French synthesis procedures (*CRDSD* 10, May 1962). But by 1966 MT research had ceased in Brussels; Euratom's involvement in MT had transferred fully to CETIS at Ispra (ch.11.1). However, by this time, the Brussels group was already mainly involved in the development of the computer-based multilingual dictionary DICAUTOM for the European Coal and Steel Community (later, as EURODICAUTOM (ch.17.7), expanded for the European Community.

5. 7: Research in West Germany (1963-68)

In West Germany MT activity before the mid-1960's was surprisingly limited. There was some research at Freiburg under Herbert Pilch on syntactic analysis and synthesis of English, intended for a MT system of the 'syntactic transfer' type, possibly with Finnish as a target language (Zint 1967). At the University of Cologne, Paul O. Samuelsdorff worked for a time on Hebrew-English MT, basically a 'direct' system (Zint 1967); subsequently Samuelsdorff experimented with an English-German system on similar principles (Bruderer 1978:129-131)

A more substantial experimental system in West Germany was the investigation at IBM Deutschland (based in Stuttgart), from 1963 until about 1968, of a system for English-German MT of texts in the fields of data processing and electronics (Schirmer 1969, Batori 1969). The system belonged clearly to the 'direct translation' type: analysis and synthesis were closely intertwined ("eng miteinander verflochten"); and the basis of its syntactic method was similar to Garvin's fulcrum approach (ch.4.6). Initial dictionary lookup (a bilingual English-German dictionary of 11,000 entries) was followed by programs for homograph resolution, identification of coordinate structures and constituent structure relationships, modification of certain English syntactic structures into ones corresponding to German usage (e.g. present participles into subordinate clauses), generation of German noun, adjective and verb endings, and transformation of structures into German word orders. The research at IBM in Stuttgart did not, however, progress beyond a tentative experimental stage.

5. 8: Other West European groups.

As in the United States there were a number of short-lived or abortive groups. In Scandinavia a group was founded at Stockholm in November 1960, led by Stig Comet, with the intention of establishing a MT society. In Finland a group was set up at Kasvatusopillinen Korkeakoulu, Jyväskylä, led by Lennart Schäring and Teuvo Kuikka, to begin preliminary studies for Finnish-Swedish MT (*CRDSD* 11, Nov 1962). Finally, a group was set up under Nicola Wolkenstein at the University of Pisa in 1964, planned to develop an Italian-English MT system (*CRDSD* 13, Nov 1964). Nothing was heard subsequently of any of these projects.