Chapter 11: Other indirect systems, 1965-1975

11. 1: EURATOM, European Scientific Data Processing Centre(CETIS), Ispra (Italy) (1967-75)

When EURATOM (the European Atomic Energy Community) acquired the Georgetown Russian-English system in 1963 (Ch.4.3) the primary purpose was to provide scientists at Ispra and other EURATOM research centres with a rapid and economic translation service. Translations were delivered without post-editing and, although it was accepted that the quality was not high, for the purposes of 'current awareness' and quick information users were satisfied to receive some form of translation rather than none at all (Perschke 1968).

The secondary purpose of the acquisition was to use it as the basis for research at the European Scientific Data Processing Centre, CETIS (Centre Européen pour le Traitement de l'Information Scientifique), which had been set up by EURATOM at Ispra in 1959. The Centre came under the aegis of the Commission of the European Communities in 1967, when the European Communities were formed from a merger of EURATOM, the European Coal and Steel Community and the European Economic Community. The research at CETIS was conducted under the direction of Sergei Perschke. The first step was to extend the application into the field of documentation by the development of a system for the automatic indexing and abstracting of Russian documents (Perschke 1969). Its software was based on the SLC-II macro-assembler programming system, an improved version of the system developed by A.F.R.Brown to run the Georgetown (GAT) translation programs (Ch.4.3) and after subsequent enhancements and improvements it became from 1972 the fully automatic information retrieval system FAIRS (Perschke et al. 1974)

Support had earlier been given to the development at the University of Brussels of a Russian-French system, latterly based on the Georgetown system (Ch.5.6) Work began at CETIS in 1967 on a Russian-English system, now based on SLC-II, and designed for the IBM 360/65. The principle problems with the Georgetown system were seen as the poor handling of homographs and the lack of full syntactic analysis. In the CETIS system SL analysis was to produce a full 'surface syntactic' representation with some semantic markers. The syntactic model adopted was Ceccato's correlational analysis (Ch.5.3). Perschke had been a member of the Milan group before going to CETIS, and Ceccato's research had received funds from EURATOM in earlier years (Ceccato 1967) In the spirit of Ceccato, Perschke (1970) included as a long-term objective the introduction of semantic analysis of lexical items and of "a system of associations, i.e. the formal description of general knowledge about the relation of things to each other." More immediately, the CETIS project was to be the basis for a multilingual system, designed to cope with the pressing needs of the European Communities translation services (ch.14). Although research would begin with Russian-English the intention was to add other TLs soon afterwards (Perschke 1970a). The system was designed therefore on the 'transfer' model, with separation of SL analysis and TL synthesis. In fact, the system was modelled on the 'syntactic transfer' approach advocated earlier by Yngve (Ch.4.7), initially operating "at a very low level of semantic analysis" (Perschke 1968), but capable of progressive refinement "because of its open-ended design" (Perschke 1970a). There were five stages: Pre-editing (as in the Georgetown system), Dictionary lookup and morphological analysis (rather similar to the Systran approach), Transfer (which combined SL syntactic analysis and SL-TL converstion into TL tree representations), and Morphological synthesis (Perschke et al. 1974) However, only the analysis programs were developed in any detail as most activity concentrated on the information retrieval applications, and so the full system did not reach a prototype implementation.

In effect, the CETIS research was overtaken by events. For its practical Russian-English MT needs, EURATOM installed Systran at Ispra in 1970 (Ch.12.1). By 1975 there were far more

advanced MT projects under way within the European Communities, the Leibniz group was laying plans for collaborative research on a multilingual system and there were beginning to emerge ideas for a European multilingual system based on the most advanced linguistic and computational developments (Ch.14)

11. 2: University of California, Berkeley (1967-75)

Research on Chinese-English MT began at Berkeley while Lamb was still director, under C.Y.Dougherty (Ch.4.10) Around 1967 there was the beginning of a certain amount of cooperative work with the centres at Bunker-Ramo and the University of Texas (Wang et al. 1973) Berkeley was to be responsible for a large machine-readable dictionary and the development of an automatic parser and grammar; Texas (Ch.4.11) joined in the lexicographic work and Bunker-Ramo (Ch.4.6) undertook to apply its fulcrum technique in the development of "interlingual mapping rules for Chinese to English" (taking as input the structural trees provided by the Berkeley parser), and also the development of an English synthesis program (cf. Chai 1968). However, in 1967 MT research at Bunker-Ramo ceased, and since the Berkeley parser was also only at an initial stage of experimentation, "no fruitful results came out of the cooperative venture."

In 1968, US research on Chinese-English MT was consolidated at Berkeley, under the direction of William S.Y.Wang who had previously been on a Chinese MT project at the Ohio State University (Ch.4.13). Closely linked to the MT research was the Dictionary on Computer (DOC) project concerned primarily with Chinese historical phonology. The MT project itself, the Project on Linguistic Analysis (POLA), continued the earlier emphasis at Berkeley on the problems of analysing and parsing Chinese. There was also considerable research on a Chinese-English dictionary (70,000 entries in 1973) and on comparative analyses of Chinese and English syntax, but almost no work was done on programs for the synthesis of English output. Details of POLA are to be found in Wang et al. 1973, 1974, 1975, Wang 1976).

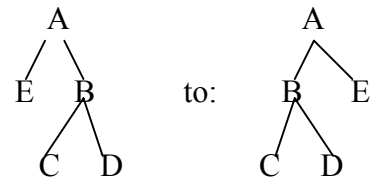
The POLA system QUINCE was based on the 'transfer' approach: "Chinese sentences constitute the source language input. This is submitted to the parser and analyzed into structural trees. Interlingual processes then apply to these structures to map them into the appropriate English structure. These structures are then used for synthesis into the target English output by applying the necessary surface structure rules" (Wang et al. 1973) The translation process had six stages: Input, Segmentation, Lexicon, Grammars, Transfer, Extract.

The input of Chinese text characters causes problems for all MT projects: POLA experimented with a number of systems for coding into Standard Chinese Telegraphic Code (including the one developed at IBM and the Itek Corporation, Ch.4.2, but it was found generally unreliable and difficult to operate) A large part of the POLA research effort was devoted to this task before the project ended in 1975.

The next stage after input (Segmentation) dealt with one of the problems peculiar to Chinese, namely the lack of sentence-delimiting and word-delimiting symbols. Using punctuation marks, prepositions and conjunctions it derived a preliminary segmentation of Chinese texts into tentative sentences and 'subsentences'. The stage of dictionary lookup (Lexicon) could then search for the longest matching sequences of Chinese characters. By 1975 the Chinese-English dictionary comprised over 82,000 entries (30% general vocabulary, 60% physics, 10% chemistry).

The parsing of Chinese ('Grammars') was performed by a battery of five 'subgrammars', which acted on tree and subtree structures. Syntactic analysis of Chinese is made difficult by the absence in Chinese of surface markers for tenses, gender, cases, and grammatical roles. With most Chinese words having multiple syntactic functions, it was necessary to include semantic markers in dictionary entries which could be referred to during analysis for homograph disambiguation.

In the next stage, Transfer, the synthesis of English-like structures was begun. Structural transfer rules were formalised as generalised tree transducers (cf. Ch.9.14), e.g. converting:



For example, these rules (applied recursively) could transform the Chinese structure:

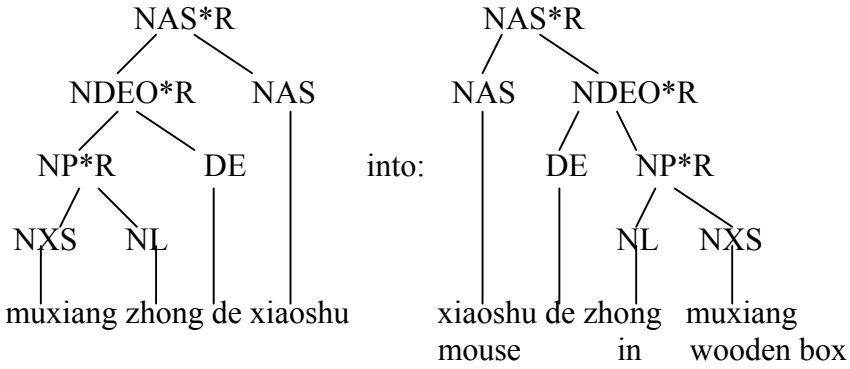


Fig.20 – POLA structural transfer

Subsequent parts of Transfer would substitute TL (English) lexical items, delete *de*, insert *the*, and (since the grammar code NAS indicates a plural noun) change *mouse* to *mice*. In the final stage ('Extract'), the English sentence string would be stripped from the TL tree structure, and appropriate morphological adjustments performed for English output.

On the computational side, QUINCE had a number of features in comon with other experimental systems at the time (in particular CETA and LRC, Ch.10.1 and 10.2). It followed the sound practice of distinguishing clearly between linguistic data and algorithmic procedures, and programming complexity was avoided. At an earlier phase of the project, it had become apparent that the monolithic grammars were growing unmanageable, and in 1973 the system was redesigned on 'modular' lines (reaching final implementation in May 1974.) The grammar was divided into a flexible sequence of 'subgrammars' (formulated as phrase-structure gramars) allowing linguists to rewrite small areas of the grammatical program without fear of impairing the system as a whole. It was intended that some of these subgrammars could incorporate heuristic techniques (presumably on lines indicated by Garvin, Ch.4.6 above), although at the close of the project in 1975 none had in fact been implemented. There was also some interest in introducing intersentential analysis, particularly since written Chinese itself does not indicate sentence boundaries. Finally, The POLA project attempted to make the system 'portable'. The whole program was written in a 'structural programming language' (GASP), apparently with similarities to PASCAL, which it was claimed (or hoped) made the system "highly machine independent" and easily adaptable to any large-scale third generation computer (Wang et al. 1975)

11. 3: Other US projects

Bruderer (1978) reports a project by Nitaya Kanchanawan of the Ramkhamkaeng University of Bangkonk on an experimental Thai-English MT system. Research began at the Florida State University, Tallahassee, in 1974 and transferred to Texas in 1975. It was a strictly limited system on simple intransistive sentences, with a dictionary of just 345 Thai entries. Apparently, the parser combined a Backus-Naur formalism of a context-free grammar with transformational rules. The analysis was oriented towards TL (English) structures producing 'deep structures' for TL synthesis; however, there was no semantic analysis and therefore no resolution of

polysemes. The MT design would appear to have had little or no connection with the major LRC project at the time (Ch.10.3)

Equally tentative was the research at the Dropsie College for Hebrew and Cognate Learning, Philadelphia, by James D. Price. In 1969, he submitted a doctorate dissertation which described work on a context-sensitive phrase-structure grammar of Hebrew in preparation for a projected Hebrew-English MT program, apparently on the 'syntactic transfer' model (Price 1969)

11. 4: University of Münster (1966-1969)

The research by Klaus Brockhaus (1971) at the University of Münster was devoted to the design of a small-scale system for reciprocal translation of English and German. A primary objective of this research, completed in 1969, was the development of methods of phrase-structure analysis for English and German and of generalised algorithms for syntactic transfer. The system was tested on a very small corpus of deliberately limited nature. The partial syntaxes for German and English did not, for example, treat adverbs or conjunctions, or the English progressive present form of verbs; and the partial vocabularies excluded all instances of homography. The aim was solely to test the adequacy of the syntax program; problems of semantics were deliberately ignored. Although obviously limited, the work represents an early attempt to generalise syntactic transfer algorithms in the form of abstract tree-transduction rules (Ch.9.14) Brockhaus moved subsequently to Heidelberg where he directed the SALAT project (Ch.16.1)

11. 4: Bulgarian Academy of Sciences (1964-76)

Research in Bulgaria was stimulated by preparations for the 5th International Congress of Slavists held in Sofia in 1963 (Ljudskanov 1966). At this time began studies by Leskov on problems of applied linguistics with reference to MT between Slavic languages; and the theoretical research by Alexander Ljudskanov on Russian-Bulagarian MT.[1]

In 1964 the section 'Automatic translation and mathematical linguistics' was established in the Mathematical Institute of the Bulgarian Academy of Sciences, Sofia. Its aims were the compilation of a Russian-Bulgarian dictionary for mathematics, the construction of a general algorithm for lexical analysis, and the commencement of work on a Russian-Bulgarian MT system, initially a word-for-word and dictionary system.

Ljudskanov's main contributions were his theoretical studies of human translation and MT strategies (Ljudskanov 1972). One of his arguments was that translation does not necessarily require full ('deep') understanding of the subject matter being translated. What is required is knowledge of how to select the appropraite TL expressions for a given SL text. Even in cases where appraentlty the translator refers to extra-linguistic knowledge in order to understand the text, "the referential approach is aimed not at establishing information about the real world for its own sake but at using the real world to establish information about the corresponding devices of the language" (Ljudskanov 1968). He introduced the idea of 'necessary translation information' consisting of the basic lexical information and the additional contextual information necessary for interpretation. The problem for MT is to determine what this information is to be in particular cases. The contextual information will vary according to the language pairs involved; it is not essential always to analyse to the 'deepest' semantic levels; MT can work sometimes at more superficial levels. His conclusion was that MT requires a different type of model than the general linguistic models developed by CETA and by Mel'chuk, for example. His inclination was, therefore, away from abstract interlingual approaches and towards a practical 'transfer' approach.

[1] An appraisal of Ljudskanov and the MT research in Bulgaria is: E. Paskaleva 'Alexander Ljudskanov', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J. Hutchins (Amsterdam: John Benjamins, 2000), 361-376.

Some results were achieved after 1970 on an experimental Russian-Bulgarian system for mathematics texts, implemented on a Minsk-32 machine. The system was apparently of the 'syntactic transfer' type, based on dependency grammar. A basic assumption was the isomorphism of separate linguistic levels, permitting standard forms of analysis at morphological, syntactic and semantic levels (Zarechnak 1979: 63). Certain parts of analysis and synthesis programs were, therefore, to be worked out as 'universal' algorithms, i.e. much as tree transducers may be regarded. It is unclear how much of the system was programmed; a large part of the research programme was devoted to quantitative and statistical studies of Bulgarian. The project appears to have ended before Ljudskanov's death in 1976 (Bruderer 1978).

11. 5: Projects and research in the Soviet Union (1966-76)

Reactions to the ALPAC report by MT researchers in the Soviet Union were highly critical. In 1969 Kulagina and a number of colleagues wrote (Locke 1975): We wish to declare decisively that this view has no real support: it is founded upon a failure to understand the problem in principle and confusion of its theoretical, scientific and practical aspects. The fact that machine translation has been ineffectual in practice to the present should, in our opinion, lead to an increase rather than a decrease in efforts in this area, especially in exploratory and experimental work

Evidently, Soviet researchers were coming under as much pressure as their US colleagues to deliver working systems. As in the United States, those responsible for funding research began to lose interest and an era of lower MT support began (Roberts & Zarechnak 1974). There is then evidence of some decrease in Soviet research activity after the mid 1960s, but it is clear that MT did not suffer the virtual eclipse that occurred in the United States: research continued on both the theoretical and the practical fronts. The systems of this period were based on both 'indirect' strategies; the more theoretical being interlingual, the practical systems being transfer.

The most important example of advanced theoretical MT research was the work of Igor A. Mel'chuk on the 'meaning-text' stratificational model, which has been described above (Ch.10.2). Research on this approach was pursued by workers at a number of Moscow research institutes until Mel'chuk was dismissed from his post for political reasons in 1976 (Mel'chuk 1981; *Survey* 23(2), 1978, p.126-140)[2]. Mel'chuk emigrated to Canada, where since 1977 he has continued his research in the Department of Linguistics of the University of Montreal; two years later his co-worker Alexander Zhol'kovsky, emigrated also. His other principal collaborator, Yurii Apresyan, has apparently been able to continue some experimental work on a French-Russian system based on the 'meaning-text' approach at Informelektro, the information centre for the Institute of Electrotechnology in Moscow (Kulagina 1976, Bruderer 1978, Marchuk 1984). MT activity at the Moscow State Pedagogical Institute for Foreign Languages, where the Laboratory for Machine Translation had latterly been the principal focus for work on the English-Russian 'meaning-text' model, was limited by 1977 to partial algorithmic simulations, according to Marchuk (Bruderer 1978: 158) Olga Kulagina, another of Mel'chuk's principal collaborators, was evidently able to remain active on the development of the French-Russian system at the Institute of Applied Mathematics (Ch.6.2), until research on this 'transfer' system was transferred to the Centre for Translation (see Ch.18.1 below) Other researchers of the group have apparently dispersed, with some able to maintain an interest in MT (e.g. Shalyapina 1980).

In the late 1960s and early 1970s there was research at a number of other Soviet institutions. By 1969 an operational English-Russian system had been developed at the Central Scientific Research Institute for Patent Information and Techno-Economic Research (Tsentral'nyi nauchno-issledovatel'nyi institut patentnoi informatsii) by A.L.Vasilevskii, L.G.Kravets, V.A.Moskovich, G.A.Tarasova and others (Kulagina 1976). Research on the system for translating patents had

[2] See also: I.A.Mel'čuk 'Machine translation and formal linguistics in the USSR', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J. Hutchins (Amsterdam: John Benjamins, 2000), 205-226.

began in early 1963, and the first experimental version was tested between 1964 and 1966, initially on the 'Strela-3' and later on the 'Ural-4' (Kravets et al. 1967). Development on later versions continued until 1969, when the operational system was implemented (Vasilevskii et al. 1971). The Moscow Patent Office system was a bilingual 'transfer' system designed specifically for translating the artificially structured texts of patents. Built on the 'syntactic transfer' approach, English texts were analyzed as dependency structures (using methods developed earlier by Mel'chuk); in the Transfer stage, dependency trees were converted into equivalent Russian 'surface' structures and Russian lexical units substituted, involving familiar methods of distinguishing homographs and resolving problems of multiple equivalents by reference to syntactic contexts and semantic compatibilities (using semantic features such as 'concrete', 'abstract' etc.); finally the stage of Russian synthesis produced morphologically and syntactically correct output. However, it appears that a substantial burden was carried by dictionary information rather than grammar routines (Lawson 1983), thus the system may have been more like 'direct translation' systems than a true 'transfer' type. The first dictionary was compiled for texts on internal combustion engines (Vasilevskii et al. 1971). Development continued until about 1977 according to Marchuk (Bruderer 1978: 158), and it was also installed at Atominform, the Information Centre for Nuclear Energy, Moscow, under V.M.Kalinin.

Research in Leningrad continued during the 1960s and early 1970s at the Laboratory of Mathematical Linguistics on an English-Russian system under S.Y.Fitialov, G.S.Tseitin and B.M.Leikina (Ljudskanov 1972: 187; Bruderer 1978). Work on the experimental 'syntactic transfer' system (with dependency structure analysis) began in 1966, but apparently progressed very slowly; Kulagina (1976) reported that only a very limited number of trial translations had been completed by 1976, and it would seem that research ended shortly afterwards. Evidently the interlingual research led by Andreev (Ch.6.4) had already ceased in the late 1960s; although Andreev himself has continued to contribute to MT theory.