Chapter 6: Groups and projects in the Soviet Union and Eastern Europe (1955-1967)

Research in the USSR on MT began soon after the demonstration of the Georgetown-IBM experiment in early 1954 (ch.2.6). Three major research groups started at this time; two in Moscow and one in Leningrad. One Moscow group was set up at the Institute for Precision Mechanics and Computer Technology; another was established in the Steklov Mathematical Institute. The third Moscow group was formed at the Electromodelling Laboratory of the All-Union Institute of Scientific and Technical Information. The Leningrad group was set up as the Experimental Laboratory for Machine Translation in the University of Leningrad. Shortly afterwards, a number of other groups with MT research interests were formed: one at the First Moscow State Pedagogical Institute of Foreign Languages, and another at the Institute of Linguistics in Moscow. Other groups began at Gorky, Kharkov, Kiev, Petrozavodsk, Tiflis, and Yerevan. The growth in MT research workers was very rapid, from a handful experimenting in 1954 on English-Russian and French-Russian to many hundreds by 1959. A conference on MT held in Moscow in May 1958 was attended by 340 representatives from 79 institutions (Harper 1961) Many had no doubt only a passing interest in the field, but nevertheless the amount of activity was impressive. From the beginning, Russian researchers were particularly anxious the follow the progress of MT in other countries, and many of their publications summarize and comment upon current developments; particular mention should be made of the invaluable manual and bibliography by Mel'chuk & Ravich (1967).

6. 1: Institute of Precision Mechanics and Computer Technology (ITMVT)

Research on an English-Russian system began in January 1955 at the Institute of Precision Mechanics and Computer Technology (Institut tochnoi mekhaniki i vychislitel'noi tekhniki AN SSSR); by the end of 1955 the first experimental tests were done on the BESM computer of the USSR Academy of Sciences.

The experiment was described in detail by Mukhin (1956) and Panov (1960). The aim was similar to that of the Georgetown-IBM experiment: to demonstrate the technical feasibility of MT. A dictionary of 952 English and 1073 Russian words was compiled to translate a mathematical text on differential equations. The simplistic adhocness of the approach can be illustrated by the procedure for translating *much* and *many* (Panov 1960: 27-28; Panov 1960a), a sequence of yes/no questions linked by transition codes.

1(2,3)	Check immediately preceding word for *how*
2(0)	*skol'ko* (numeral, invariable)
3(4,5)	Check immediately preceding word for *as*
4(0)	*stol'ko zhe* (numeral, variable)
5(7,9)	Check given word for *much*
6(0)	Not to be translated (adverb)
7(1,11)	Check immediately preceding word for *very*
8(0)	*mnogii* (adjective, hard stem, with sibilant)
9(8,12)	Check preceding word for preposition, and following word for noun
10(0)	*mnogo* (adverb)
11(12,10)	Check following word for noun
12(0)	*mnogo* (numeral, variable)

(where the two numbers in brackets indicate the next rule if the answer is ‘yes’ or ‘no’ respectively.). To translate *The subject would have been much better standardized*, the sequence is: 1 - answer ‘no’, go to 3; 3 – ‘no’; 5 – ‘yes’; 7 – ‘no’; 11 – ‘no’; 12 - translation is *mnogo*. In *This is*

most useful and for many reasons the sequence is: 1,3,5,9,8 giving *mnogii*. The computational (and linguistic) cumbersomeness of such procedures is on a par with those of the Georgetown-IBM experiment (ch.4.3; cf. Panov (1960) and Sheridan (1955))

Experiments were also made with texts from *The Times* newspaper and from Dickens' novel *David Copperfield*, without any expansion of the dictionary. The Dickens translation was admitted to be "extremely imperfect" but it was believed to show that although designed for mathematics sufficient general vocabulary had been incorporated to give reasonable results (Panov 1960).

Just as Dostert had at Georgetown (ch.4.3), Panov (1956) drew certain principles for future MT work from the experiment. He advocated: the maximum separation of the dictionary from the translation program; the separation of analysis programs and synthesis programs; the storage of lexical items under stem forms; the inclusion of grammatical information in dictionaries; and the determination of multiple meanings from contextual information. Similar principles were, of course, reached by the American researchers at an early stage, as we have seen.

The linguistic aspects of the program were the work of I.K.Bel'skaya, who continued the further development of the English-Russian system at ITMVT. This system (Bel'skaya 1957, 1960) was one of the best known of the early Soviet MT research efforts. Its dictionary contained 2300 words from texts on applied mathematics. The restriction to one relatively narrow scientific field was justified both to ensure a reasonable reduction of polysemy problems and to keep within the limitations of the available computer facilities. English words were entered in canonical forms, thus necessitating the conversion of words such as *wanted* to *want*, *stopped* to *stop*, *lying* to *lie*, etc. The system had three phases, each consisting of a number of cycles (or 'passes'). 'Vocabulary analysis' included dictionary search, identification of the word classes of unmatched lexical items and resolution of homographs by recognition of morphological forms or by examination of the immediate context (on the same lines as described above by Panov). The second phase was 'Grammatical analysis': first an examination of verb forms (to determine tense, mood, etc. and whether inversion of Russian word order would be required), then identification of phrase boundaries (from punctuation marks and verb groups), recognition of noun groups and then of constituent adjectives and their relationships to nouns (for ensuring correct agreement in Russian), lastly recognition of English structures which would need word order changes for Russian. The final phase was the 'Grammatical synthesis' of Russian, i.e. the inflection of Russian verbs, adjectives and nouns, in that sequence. (This last phase was considered to be sufficiently general that it could be applied to analysis output from languages other than English.)

The program worked well for those sentences on which the algorithms had been based, and Bel'skaya (1960) also reported satisfactory manually simulated tests on 100 "unknown" texts, including literary texts by Galsworthy and Edgar Allen Poe. (Illustrative extracts were included in her article.) She claimed that results on the latter, together with the earlier tests on Dickens (above), demonstrated that "the applicability of MT depends on whether it is possible to identify the implicit set of rules governing this or that particular sphere of language application, be it as narrow a sphere as say, Wordsworth's poetry, and further, whether these rules can be formulated into a formal set." Bel'skaya was one of very few MT researchers believing still at this date in the possibility of MT of poetic works.

It would appear that the 1958 version of the system was not implemented on a computer (Harper 1961) It would seem very likely that the system was beset by the same problems encountered in other empirical systems, principally the inadequacies of ad hoc routines not derived from extensive examination of text (Harper 1961) The inadequacies of the word-for-word approach of the system were well described by Rozentsveig in 1968 (quoted by Roberts & Zarechnak 1974): the programs "did not compare the correspondences between the systems of the target and source languages in translating. It is therefore significant that the synthesis of a Russian sentence was obtained depending on the English input information for each word and the position of each word within the English sentence. Synthesis routines were used after the word order in English had been

transformed in accordance with the requirements for word order in Russian." Similar criticisms were made of the word-for-word 'direct' approach in US systems (cf.ch.8 below)

The Institute's work on translation from Russian has been described by T.N.Nikolaeva (1958).[1] The intention was to develop a program for analysing Russian which could be used for any target language. The Russian dictionary contained therefore only grammatical information and no TL equivalents. Morphological analysis preceded dictionary lookup in order to determine the 'dictionary form' (e.g. for nouns, the nominative singular). Problems caused by mathematical formulae were dealt with by assigning word classes (adjective, noun) as appropriate. A good deal of attention was paid to problems of Russian *-sya* verbs (which may be reflexives or passives), and to the homography of the case endings. Syntactic analysis aimed to identify the roles of components within the sentence (e.g. 'subject', 'direct object', 'instrument' etc.); each word was examined in turn, in a left to right scan, with backward and forward consultation of adjacent words as necessary. It is evident that this project involved extensive study of Russian morphology and syntax. The analysis program was written for the KIEV computer of the USSR Academy of Sciences, and was tested in April 1960 (Mukhin 1963).

The Institute's MT research effort included work on German-Russian MT. A German-Russian glossary for which "some 1500 pages of mathematics were studied" was compiled by S.S.Belokrinitskaya, who also investigated the semantics of German prepositions (Harper 1961). The translation algorithms were constructed on the same principles as those for the English-Russian system, with an enhanced dictionary lookup program in order to deal with the complexities of German compounding, and additional components in the syntactic analysis to identify subordinate clauses (Kulagina et al. 1961).

There was also research at ITMVT on translation from Chinese and Japanese. Some early illustrations were included by Panov (1960) and by Bel'skaya (1957) to demonstrate the general applicability of the method used in the English-Russian system. In later years, research followed the lines indicated by the Russian-English system of Nikolaeva (Kulagina et al. 1961). Given the non-inflectional character of Chinese, syntactic analysis was the main focus, designed primarily to identify functions which would need to be represented in Russian by inflected forms (e.g. case endings of nouns). In the case of Japanese the problem of identifying word, phrase and sentence boundaries was given special attention.

It is clear that Rozentsveig (1958) was correct to characterize the efforts of this group as "the rapid achievement of immediate, practical results", in conjunction with "careful, detailed investigation of linguistic material, especially lexical." In many respects, ITMVT's philosophy was much like that at Georgetown (ch.4.3)

6. 2: Steklov Mathematical Institute (MIAN)

Research at the Steklov Mathematical Institute of the Academy of Sciences (Matematicheskii institut imeni V.A.Steklova AN SSSR, i.e. MIAN) started at the end of 1954 under the general direction of the mathematician and computer specialist A.A.Lyapunov. After an initial 'empirical' experiment on French-Russian MT, research at the Institute was characterised by a strong theoretical orientation.[2]

The French-Russian MT project developed by Olga S.Kulagina (in collaboration with Igor A.Mel'chuk of the Institute of Linguistics) produced its first translations on the STRELA computer in 1956. Testing continued until 1959. (Examples of translations can be found in Roberts &

[1] For more information about ITMVT and relations with Motorin and the KGB see: Marčuk, Ju.N. 'Machine translation: early years in the USSR', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 243-251

[2] See also: Kulagina, O.S. 'Pioneering MT in the Soviet Union', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 197-204; and Kulagina, O.S. *Issledovanija po mašinnomu perevodu*, Moskva: Nauka, 1979.

Zarechnak (1974) and in Papp (1966: 109).) The translation algorithm was compiled empirically; first on the basis of existing human translations of French mathematical texts, then by supplementing the initial rules by consultation of further texts and of textbook grammars of French. As a result, some of the traditional word classes were modified. The dictionary, which contained 1200 French stems representing 2300 words, was compiled from examination of 20,500 running words of texts by the mathematicians Picard, Borel and Appel (Papp 1966, Harper 1961) The stem form was identical with the singular for most nouns and with the masculine singular for most adjectives. Some verbs were listed under several stems (e.g. *faire* under *fais*, *fai*, *fass*, *fe*); irregular verbs had multiple entries, often full forms (e.g. *est*, *sont*, *soit*, *soient*). Excluded were verb forms which would not be expected in mathematical texts, such as first person singular and second person singular forms. Idiomatic constructions were included as units, e.g. *mettre en doute* ('call into question'). Also included as 'idioms' were compound conjunctions and adverbs such as *le long de* ('along'), *parce que* ('because') and *à peu près* ('approximately'). Dictionary entries included grammatical information, not only genders, plural type, etc. but also which prepositions may precede or follow, providing clues for the selection of Russian case endings. In general only one Russian equivalent was provided, on the grounds that the mathematical context reduced problems of homography. The routines for syntactic analysis consisted of complex series of yes/no questions (as in the English-Russian system of Panov and Bel'skaya). There seems to have been no attempt to describe the syntactic functions of the French homographs independently of the requirements of Russian syntactic and lexical formations (Harper 1963)

One by-product of this research was the advancement of programming theory, i.e. the identification of elementary algorithmic operations and work towards the development of programming languages (Rozentsveig 1958); the inadequacies of the existing languages for non-numerical applications were as obvious to the Russians as they were to the Americans. Another was the development of Kulagina's set theory model of language, in which grammatical categories and relationships can be defined on formal mathematical foundations. The definitions produced enable analysis of sentence elements into dependency-like structures and semantic components. For example, "the words *thick book* in the sentence *the thick book lies on the table* can be reduced to the element *book* or can be replaced by the element *thing* or the element *it"* (Rozentsveig 1958). Kulagina's model opened the strong vein of mathematical linguistics in the Soviet Union. In Oettinger's (1958) view Kulagina's switch away from practical MT to theoretical issues was a consequence of encountering problems when the techniques of the French-Russian system were applied on a larger scale. Whether this was the reason or not, henceforth the research strategy at MIAN could be characterised (Rozentsveig 1958) as "the effective practical realization of machine translation only as the result of profound theoretical research in the area of mathematics and linguistics." In this regard, Kulagina's work on mathematical linguistics was considered typical.

Improved versions of the algorithms were incorporated in the English-Russian MT at the same Institute (MIAN), developed under the direction of T.N. Moloshnaya.[3] After the experience of the empirical approach to French-Russian MT, a more formal methodology was adopted (Rozentsveig 1958). For syntactic analysis a dependency model was adopted for the identification of noun and verb phrase (syntagm) structures. To cope with problems of morphology Moloshnaya found it necessary to establish a new system of word classes for the two languages. From a distributional analysis based on methods derived from Fries (1952) and Jespersen (1937), Moloshnaya identified 19 word classes for English and 17 for Russian. Homographs were resolved by a routine based on the examination of the possible word classes of contiguous words. The homograph routine preceded the dependency analysis of phrase structures. The experiment was on a relatively small scale, the dictionary containing just 1026 items from a text on differential equations (Papp 1966). Some steps were taken towards a multilingual system by dividing the

[3] See also: T.N.Mološnaja 'My memoirs of MT in the Soviet Union', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 227-231

dictionary so that the Russian part could be used with other languages and by making English analysis and Russian morphological synthesis independent programs. (Rozentsveig 1958, Kulagina et al.1961)

6. 3: Institute of Linguistics, Moscow

The Institute of Linguistics in Moscow (Institut Yazykoznaniya AN SSSR) set up a group under the direction of A.A.Reformatskii. Its involvement was primarily at a theoretical level, working often in collaboration with other research groups. As we have seen, Igor A.Mel'chuk collaborated with Kulagina on the MIAN French-Russian system. Following this, he investigated requirements for a Hungarian-Russian MT system[4]. Hungarian was chosen as presenting within itself many of the special difficulties met with in a number of languages, e.g. as an agglutinating language it posed problems common to Turkic languages spoken in the Soviet Union, its compound nouns and separable verb particles had similarities to features in English and German, and its word order was completely different from that of Russian. If a satisfactory approach could be found for Hungarian, then this might provide clues for solving similar problems in other languages (Papp 1966). Mel'chuk devised rules for morphological analysis, dictionary searching, homograph resolution, recognition of sentence structure, and Russian sentence synthesis. Although a selective glossary was compiled, the study was concentrated on the investigation of algorithmic problems. As a consequence of this research on Hungarian, Mel'chuk came to formulate his notion of an interlingua.[5] The problems of Hungarian word order compelled the abandonment of a word-for-word approach (which might be feasible for French-Russian) and favoured investigation of common syntagmatic structures (e.g. of possession, adjectival modification). A similar investigation for other language pairs would build up a series of syntactic configurations, some common to all languages but most common to only some. From this set of interlingual structures would be selected those needed for particular SL and TL texts. A similar analysis of lexical differences and equivalences would produce sets of interlingual semantic units (e.g. indicating comparison, negation, 'larger than normal' size). In this view, the interlingua is the sum of all correspondences of the languages involved. The subsequent development of Mel'chuk's ideas resulted in his well-known 'meaning-text' model of language and in the elaboration of a highly influential MT model (ch.10.2 below)

6. 4: Leningrad University (ELMP)

Research at the Experimental Laboratory of Machine Translation at Leningrad University (Eksperimental'naya Laboratoriya Mashinnogo Perevoda Leningradskogo Universiteta imeni A.A.Zhdanova, i.e. ELMP), which started in 1958, was led by Nikolai D. Andreev[6]; by 1961 over 100 researchers were involved in the group (Andreev 1961). The main thrust of this MT project was the development of the theoretical basis for interlingual MT. Andreev outlined his ideas on the interlingua on many occasions (e.g. Andreev 1958, 1967) His conception of an interlingua[7] was that of an artificial language complete in itself with its own morphology and syntax (Papp 1966), and also capable of serving on its own as an 'informational language'. Decisions about the inclusion of particular features were to be based on the averaging of phenomena of various

[4] For his experiences see: Mel'čuk, I.A. '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), 205-231.

[5] For Russian conceptions of interlingua in this period see: Archimbault, S. and Léon, J. 'La langue intermédiaire dans la traduction automatique en URSS (1954-1960): filiations et modèles', *Histoire Épistémologie Langage* 19(2), 1997, 105-132.

[6] For Andreev and the ELMP see: Piotrovskij, R.G. 'MT in the former USSR and the Newly Independent States (NIS): prehistory, romantic era, prosaic time', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 233-242.

[7] See footnote 5.

languages with a weighting given to the 'major' languages manifesting those features. For example, if more of the 'major' languages placed adjectives before nouns and subjects before predicates the interlingual syntax should do so also. A considerable research effort was dedicated to the abstract modelling of the linguistic and logico-semantic foundations of the interlingua. In addition, an impressively large number of languages were investigated. Andreev (1961) mentions research on Russian, Chinese, Czech, German, Rumanian, Vietnamese, Serbo-Croatian, English, French, Spanish, Norwegian, Arabic, Hindustani, Japanese, Indonesian, Burmese, Turkish, and Swahili. It was claimed that experimental algorithms for analysis from these languages into the interlingua were being developed, but how many advanced beyond preliminary sketches is not known. Synthesis programs for conversion from the interlingua were compiled for Russian only.

Not all research at Leningrad was concerned with interlingual investigations. Some experimental work on a practical Russian-English system was undertaken. The researchers L.I.Zasorina and T.S.Tseitin adopted the valency approach to syntactic analysis (Mukhin 1963), in which each word was examined for its potential combinations with other word-types (ch.3.4 above). Apparently, the project encountered particular problems in dealing with ambiguities of Russian case endings (Papp 1966) - as did many US projects, as we have seen. It appears that only limited tests were made involving no "more than 20 translated sentences" (Kulagina 1976)

6.5: First Moscow State Pedagogical Institute of Foreign Languages (I MGIIYa)

At the Pedagogical Institute of Foreign Languages in Moscow (Pervii Moskovskii Gosudarstvennyi Pedagogicheskii Institut Inostrannykh Yazykov), research on MT began in 1957 under the direction of I.I.Revzin; in later years the director was V.Y.Rozentsveig.[8] Research here was concentrated on general theoretical studies of semantics (e.g. the work of A.K. Zholkovskii and L.N. Iordanskaya). Particularly important was the research on semantic analysis, in which semantic representations of sentences were derived from dictionary entries of words formulated as combinations of elementary 'semantic factors' and relations, e.g. 'time', 'action', 'possession' (Zholkovskii et al. 1961). The establishment of semantic factors necessitated careful analysis of semantic fields of near synonyms, e.g. differences between *property, owner, belong* and between the various possible Russian equivalents of *appear* (viz. *poyavlyat'sya, vystupat', kazat'sya, vyyavlyat'sya*, *predstavlyat*', etc.) according to their contexts. As at the Cambridge Language Research Unit (ch.5.2 above), the emphasis was on problems of synonymy and paraphrase rather than homonymy, on subtle semantic differences rather than crude lexical transfer. Much of this work was related closely to the theoretical research of Mel'chuk on MT linguistics and was to culminate in the research on an 'interlingua' English-Russian system (ch.10.2 below). The corpus for this semantic research was primarily foreign policy texts in English, French, Spanish and Russian, with MT systems as long-term objectives. (See various articles by Zholkovskii and other members of the Institute in Rozentsveig (1974).)

6. 6: All-Union Institute of Scientific and Technical Information (VINITI)

Basic linguistic studies of English and Russian syntax and morphology were conducted at the Electromodelling Laboratory of VINITI. Research by Z.M. Volotskaya on Russian verb formation, E.V.Paducheva, T.N.Shelimova on Russian prepositions and cases, A.L. Shumilina, M.M.Langleben on Russian mathematical language was mentioned by Harper (1961) and Rozentsveig (1958). Much of this work was done with problems of information retrieval in mind, in recognition of some of the close relations between the two fields (cf. the similar views of the

[8] For memoirs of Rozencvejg see: Mel'čuk, I.A. '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; and Uspenskij, V.A. 'Serebrjanyj vek strukturnoj, prikladnoj i matematičeskoj lingvistiki v SSSR i V.Ju.Rozencvejg: kak éto načinalos' (zametki očevidca)', *Wiener Slawistischer Almanach, Sonderband 33* (1992), 119-162.

Cambridge Language Research Unit, ch.5.2 above). It would also seem that a substantial proportion of the research was undertaken in connection with Mel'chuk's work on interlingual MT; some was basically of a statistical nature.

6. 7: Other Soviet groups.

Outside the main centres in Moscow and Leningrad there were many Soviet institutions involved in MT research of some kind. As we have seen, the conference in May 1958 was attended by representatives from 79 USSR institutes; some of this interest resulted in the achievements reported from a number of institutions.

For example, there was research on MT from and into Russian at the Gorky State University (Rozentsveig 1958). It was evidently on a small scale, since according to Papp (1966) the experimental English-Russian system involved a dictionary, compiled from radio engineering texts, containing no more than 500 items. More ambitious and theoretical was the important research under A.V.Gladkii on a semantic intermediary language in connection with an experimental German-Russian system at the Mathematical Institute, Siberian Division of USSR Academy of Sciences, Novosibirsk (Mukhin 1963; *CRDSD* 10, May 1962) This group later collaborated with Mel'chuk and others in the development of the 'meaning-text' approach (ch.10.2)

The MT research in institutes of the constituent republics of the Soviet Union had to tackle the additional problems of languages such as Uzbek, Georgian, Armenian and Lithuanian which had been studied much less rigorously and systematically than English and Russian. Much effort had therefore to be devoted to preliminary linguistic research. Nevertheless, there was widespread activity. There was, for example, a project at the Vilnius State University in Lithuania (Vilniaus Valstybinis V. Kapsuko Vardo Universitetas) for a Russian-Lithuanian MT system, taking algebra textbooks as a corpus (*CRDSD* 9, Nov 1961). Mukhin (1963) mentions work on the morphological analysis of Tartar and Uzbek at the Computing Centre, State University, Tashkent, and research on Georgian at the Institute of Electronics, Automation and Telemechanics of the Georgian Academy of Sciences, in Tiblisi (Institut Elektroniki, Avtomatiki i Telemekhaniki). This research was evidently on a fairly substantial scale. The group under Archil Eliashvili (and later C.B.Choidze) developed a Russian-Georgian MT system which employed a variant of Mel'chuk's dependency analysis for Russian (*CRDSD* 8, May 1961), and then turned to translation from Georgian. By 1963 (Mukhin 1963; *CRDSD* 13, Nov 1964) research had extended to systems for translating into Russian, English and German (the latter in conjunction with an East German researcher, cf.6.13 below). The Georgian group continued research until the mid-1970's (Bruderer 1978)

Some of these institutions constructed operational MT systems, as Kulagina (1976) has reported. A system for Russian-Ukrainian MT was implemented by about 1966 at the Institute of Cybernetics in Kiev. "It was tested in three variants: dictionary translation" (i.e. strictly word-for-word), "dictionary plus morphology, and dictionary plus morphology plus elements of syntax". In the early 1960's, a system for Armenian-Russian MT was completed at the Computer Centre of the Academy of Sciences of the Armenian SSR, Yerevan. Although algorithms and grammatical information were separated, and also analysis and synthesis programs, it was still basically a 'direct' bilingual system with procedures oriented to the specific language pair (ch.3.9 above) A similar system for Russian-Armenian MT was constructed at the same centre during the mid-1960's.

6. 8: Summary of the USSR scene (1955-66)

As in the United States, MT research in the Soviet Union went through an initial stage of enthusiasm for the potentialities of the computer, fuelled often by exaggerated conceptions of the power of the new tool. Contemporary illusions about 'thinking machines', often markedly similar to those in the United States and Western Europe, have been well documented in Oettinger's review of early MT efforts in the Soviet Union (Oettinger 1958) As in the US this first phase was

characterised by all-out assaults on MT systems, attempting to construct rules on an empirical basis, in the belief or hope that improvements could be made later. By about 1960, it had been realized that more substantial theoretical research was needed in order to progress beyond essentially 'word-for-word' systems (Harper 1961)

By the mid-1960's research on MT in the Soviet Union had produced a number of fairly crude operational word-for-word systems. Prospects for their improvement did not, however, seem good; and the considerable activity on MT-related linguistic theory did not promise immediately feasible systems as yet. In broad outline, the situation was much as in the United States, and funding for MT research declined somewhat after the mid-1960's. It is possible that the ALPAC report had some influence on sponsoring agencies in the Soviet Union, as Roberts & Zarechnak (1974) suggest (cf.8.11 below). However, there was not the same dramatic hiatus in the Soviet Union and MT research continued to progress steadily, as we shall see.

6. 9: Research in Eastern Europe (1957-66)

In East Europe, MT activity began not long after the first USSR experiments. Soviet research was a great influence, although there was also considerable individual innovation. As in the Soviet Union and in many Western European projects, progress was often severely hindered by the lack of computer facilities.

6. 10: Charles University, Prague, Czechoslovakia.

A group of linguists at Charles University began the investigation of English-Czech MT in 1957. A special department was established in 1959, later split into the linguistic group of the Centre of Numerical Mathematics and the Section of Algebraic Linguistics and Machine Translation in the Department of Linguistics and Phonetics.[9] Both groups worked closely with the Research Institute of Mathematical Machines (RIMM), where the programming has been undertaken. In January 1960, the first limited experiment was performed on the Czech computer SAPO. Subsequently, a second experiment was prepared for the EPOS I computer (Konečna et al. 1966) The subject field selected was electronics, for which a MT English-Czech dictionary was compiled from a text corpus of 100,000 words (Sgall & Hajičova 1966)

The usual stages of analysis and synthesis were adopted: dictionary lookup, morphological analysis, syntactic analysis, Czech synthesis. Morphological analysis matched words against a table of regular English endings; irregular forms were stored as wholes in the dictionary; and only idioms actually in the texts were included. For syntactic analysis, the project used the method of Moloshnaya (ch.6.2 above), and later a modified version of the Harvard predictive analysis technique (Konečna et al 1966; Sgall & Hajičova 1966) The Czech synthesis program was based on the generative grammar model developed by Petr Sgall; and in this connection, considerable detailed research has been conducted on the formal linguistic description of Czech (Sgall & Hajičova 1966)

The theoretical foundation of the Czech group was a 'stratificational' approach (ch.3.10; cf. also 4.10). Four levels were recognized: graphemic, morphemic (concerned with word formation), formemic (roughly concerned with syntactic structures), and semantic (concerned with semantic relations, e.g. agency, and semantic representations) The assumption was that the semantic level approaches language universality, and might function as an 'intermediary language'. Morphological analysis processed input text as far as the morphemic level; syntactic analysis as far as the semantic level. For their conception of an interlingua, the group adopted the Andreev approach (ch.6.4); i.e. not just "a net of correspondences" but "an individual language... specified

[9] For the Czech MT group see: Kirschner, Z. 'Pioneer work in machine translation in Czechoslovakia', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), 349-360

by a generative grammar" (Konečna et al. 1966). The word order in the interlingua was to retain that of the input text, thus avoiding the need to look at intersentence relations to decide on topic-comment organization (cf. the problems encountered at CETA, ch.10.1 below)

From the mid-1960's the group expanded its activities to many areas of computational and mathematical (particularly algebraic) linguistics, while still maintaining an interest in MT from the 'stratificational' approach (primarily in the form developed by Petr Sgall). No MT experimentation as such took place until 1976, however, when MT research was resumed (see ch.13.6 below)

6. 11: Hungarian Academy of Sciences.

The MT research group at the Computing Centre of the Hungarian Academy of Sciences (Department for Theoretical Questions) was established in the summer of 1962 as a consequence of a conference on MT held in March that year (Kiefer 1964). At this conference, György Hell and György Sipöczy described their experimental word-for-word Russian-Hungarian system at the University of Technical Sciences. The first task of the small MT group at the Computing Centre was to test the algorithms (involving laborious machine coding). In September 1963, Ferenc Kiefer was appointed leader of the group. Apart from the work on Russian-Hungarian, the group also investigated English-Hungarian and German-Hungarian. The group was considerably impeded by inadequate computer facilities for MT; above all, insufficient storage capacity. Consequently, the programs did not go beyond morphological analysis. The group devoted itself therefore to primarily theoretical work in the areas of syntactic analysis and semantics (Sipöczy 1964, Abraham & Kiefer 1966) Some research on mathematical linguistics applicable to MT was conducted in cooperation with Ferenc Papp and others at the University of Debrecen (Papp 1964) By 1966 research at the Computing Centre, now led by D.Varga (Kiefer had left Hungary) had turned to the development of parsing systems for language analysis in general, influenced to some extent by Mel'chuk's approach to semantic analysis (Varga 1967)

6. 12: Projects in Yugoslavia, Poland, and Rumania.

There was considerable interest in MT at Belgrade University in Yugoslavia (*TA* 1(1) April 1960; *CRDSD* 8, May 1961; *CRDSD* 13, Nov 1964). At the Institute of Experimental Phonetics, Djordje Kostić coordinated a large team compiling dictionaries and glossaries and conducting systematic studies of Serbian grammars. The original aim for a Serbo-Croat MT system was still being entertained in 1966, but evidently such ambitions came to naught.

The story was much the same in Poland. A group was set up during 1962 in the Research Centre for Applied Linguistics in the University of Warsaw. The aim was MT, but most work was of a theoretical nature in mathematical linguistics (*CRDSD* 10, May 1962; Josselson 1971)

Rather more progress was achieved in Rumania. Research began in September 1959 under the aegis of the Rumanian Academy of Sciences, and was directed by Grigore Moisil (Josselson 1971). Most work was of a theoretical nature by Moisil (1960) on systems for translating from French and Russian into Rumanian. There was, however, partial implementation by Erika Domonkos (1962) of an English-Rumanian 'direct' system. Tests of the system were conducted in 1962, evidently on a very small vocabulary. There were no routines for resolving homographs, for dealing with pronouns or even for treating irregular English verbs.

6. 13: Research in East Germany.

The East German Academy of Sciences set up its Working Group for Mathematical and Applied Linguistics and Automatic Translation (Arbeitsstelle für mathematische und angewandte Linguistik und automatische Übersetzung) in 1960 under the leadership of Erhard Agricola. Its objective was the development of the mathematical and computational foundations of practical MT systems for translating from English and Russian into German. The group was engaged in extensive lexicographic activity, e.g. the compilation of reverse dictionaries, and on fundamental

research on mathematical linguistics. By 1963 it was ready to test its ideas on an experimental English-German system, intended as a model for a practical system (Kunze 1966). The work was limited to a small corpus of just 22 sentences, but it was felt nevertheless that the practical adequacy of the approach and its formalism had been demonstrated. From 1963 until about 1966 the group directed its principal efforts to the development of a more extended experimental Russian-German system (Nündel et al.1969). Most emphasis was placed on the algorithms for Russian analysis; it was believed that, at least in principle, the problems of German synthesis had been solved during the English-German experiment. Practical difficulties with computer facilities prevented the implementation of the system; algorithms had been formalised in detail but not programmed.

Both experimental systems were of the 'direct' bilingual, one-directional type. The main stages were described (Nündel et al. 1969) as follows: 1. lexico-morphological analysis (i.e. dictionary lookup and identification of grammatical categories), 2. idiom processing, 3. syntactic analysis (based on a dependency grammar approach) and homograph resolution, 4. selection of TL equivalents, 5. preparation of information for synthesis, 6. rearrangement (including the insertion of German articles), 7. syntactic synthesis, 8. synthesis dictionary (for German inflectional information), 9. morphological synthesis. The researchers acknowledged the absence of clearly separated stages of analysis and synthesis (stages 1 and 3 were analysis routines, stages 2, 4 and 6 'translation' and stages 5, 7, 8 and 9 synthesis), and also the limitation of the analysis procedures to structures which required treatment in German synthesis.

The methodology followed familiar patterns, e.g. alphabetical sorting of texts for dictionary searching, separation of stems and endings for Russian but not for English, a dictionary for general vocabulary and special technical microglossaries, the development of the group's own classification of verbs and adjectives, identification of fixed idioms for translation before syntactic analysis, the incorporation of syntax rules in the analysis algorithm (rather than in separate tables), and so forth. The deficiencies and limitations of the 'direct' approach were readily acknowledged by the end of the project, and the practical achievements were meagre, no working system having been implemented; nevertheless there were substantial contributions by the group. These were primarily at a theoretical level, e.g. in the elaboration and rigorous formalisation of the dependency grammar model (cf. Kunze 1975) and in the investigation of syntactic ambiguity (Agricola 1967). As in many other groups, semantic problems were considered to be the greatest challenge for theoretical linguistic studies, but its members made only minimal contributions in this direction.

While this was clearly the principal East German MT project, it was not the only one. An intriguing example is the short-lived exploration at Jena by Fähnrich (1970) of a Georgian-German system; it was to be based on the MT research at the Georgian Academy of Sciences (ch.6.7 above). The argument for translating from Georgian was that numerous publications of scientific importance, in mathematics, mining, agriculture and medicine, were being neglected because they were not accessible even in Russian summaries. A short text was selected for a manual simulation of a binary 'direct' system; Fähnrich's paper made no substantial contribution, it is of interest only as one of the few attempts to tackle the problems of a Caucasian language.