Chapter 18: Projects and systems in the Soviet Union and Japan (1974-)

18. 1: The Soviet research (1974-

In the mid-1970s Soviet MT activity saw changes in both approach and personnel. The theoretical and experimental work at Moscow by Mel'chuk, Kulagina and their colleagues, and at Leningrad by Andreev, Fitialov and others (Ch.10.2 and 11.5) has been largely replaced by the more practical pragmatic activities of Piotrovskii in Leningrad and Marchuk at the Centre for Translation in Moscow.

18. 2: Leningrad, Group for Speech Statistics

Since the mid-1970s the main centre for MT research in Leningrad has been at the Group for Speech Statistics (Statistika rechi), under Raimund Piotrovskii of the A.I.Herzen State Pedagogical Institute.[1] The emphasis of Piotrovskii's research has been on stochastic modelling of languages, on the argument that natural languages are "open, dynamic and fuzzy" systems and that "MT problems cannot be treated exclusively on the basis of set theory and generative grammar" (Piotrovskii 1980). The system is being designed to be modular and "extendable without reprogramming the whole system" and capable of development "on the basis of step increments". At the centre of the system is the "automatic dictionary which is included into a linguistic data bank where information about the relationships between various linguistic and encyclopaedic objects is stored in the form of a semantic network". So far only the dictionary has been developed, and it is being used for word-for-word translations of English and Japanese patents. Later research is to develop modules for resolving lexical and syntactic ambiguity "based on an analysis of their contextual environment and the thesaurus reading" and for syntactic and semantic analysis based on dependency models and case frame grammars. In basic strategy, the system recalls the earliest suggestions of Kaplan and Harper (Ch.2.4.1 and Ch.4.4); only the mention of case frames and encyclopaedic knowledge reveals the influence of more recent developments in natural language processing.

18. 3: Moscow Centre for Translation

In Moscow the main centre for MT activity has since 1974 been the All-Union Centre for Translation of Scientific and Technical Literature and Documentation (Vsesoyuznyi tsentr perevodov nauchno-tekhnicheskoi literatury i dokumentatsii), Moscow, under the direction of Yurii N. Marchuk.[2] Systems have been developed for translating from English, French and German into Russian.

Initially there were two English-Russian projects. The first, "based on cyclic methods of analysis" and using disambiguation routines on the basis of a "contextological dictionary" (Marchuk 1977), had apparently been developed originally for translating political texts (Kulagina 1976). This system is now known as AMPAR (Avtomatizirovannyi Mashinnyi Perevod s Angliiskogo yazyka na Russkii), and it came into operational use in 1979 (Kotov et al. 1983).[3] The

---

[1] Fuller information about the project can be found in Z. Piotrowska, R. Piotrowskii, Y. Romanov, N. Zaitseva, and M. Blekhman, 'Machine translation in the former Soviet Union and in the new Russia', *International Journal of Translation* 13 (1/2), 2001, pp.87-104; and in R.G. Piotrovskij, 'MT in the former USSR and in the Newly Independent States (NIS): pre-history, Romantic era, prosaic time', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J.Hutchins (Amsterdam: John Benjamins, 2000), pp.233-242.

[2] For further information about the foundation of the Centre see: J.N.Marchuk, '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), pp.243-251.

[3] The AMPAR system had been developed by the KGB under Yurii Motorin, initially based on work by I.K.Bel'skaya at the Institute of Precision Mechanics and Computer Technology (see Ch.6.1). See Marchuk reference in footnote 2.

other English-Russian system, using a "continuous analysis method based on tree-representation of English syntax" (Marchuk 1977), would appear to have been a product of Martem'yanov's research under Rozentsveig at the Moscow Institute for Foreign Languages (Martem'yanov 1977); however it seems that it must have been abandoned shortly afterwards. The German-English system followed the AMPAR model of "cyclic-type analysis" (Marchuk 1977); this system (NERPA) became operational in 1981 (Kotov et al. 1983). The system for French-Russian translation (FRAP), described by Marchuk (1977) as employing a "continuous filter-type analysis", has been developed under Leont'eva. This system also became operational in 1981 (Kotov et al. 1983).

18. 4: The AMPAR system

The AMPAR system is explicitly characterised as a one-directional bilingual 'direct translation' system. Marchuk et al. (1975) point out the integration of linguistic data and algorithmic processing and stress, in particular, the interlocking of analysis and synthesis procedures. Further indications of its 'direct translation' character (Ch.3.9) are the lack of specific stages of syntactic analysis and synthesis: the syntactic rearrangement of SL text sequences into appropriate TL sequences is determined by local lexical and grammatical rules.

The AMPAR program has 17 stages distributed in the following phases (Marchuk et al. 1975, Marchuk 1984):

(1) search of the basic dictionary, and morphological analysis

(2) identification of segments (by punctuation marks), and assignment of tentative grammatical categories to words not found in the dictionary

(3) search of the idiom dictionary

(4) treatment of English *-ing* forms (some being left to later phases)

(5) homograph resolution on basis of grammatical categories

(6) 'grammatical analysis' determining the information needed for synthesis (e.g. regarding Russian aspect), establishing the syntactic functions of words on the basis of morphological, syntactic and semantic context, rearranging English negative constructions to conform to Russian negative forms and analysing residual *-ing* forms

(7) translation of 'unambiguous' words (i.e. those English words which have only one Russian equivalent)

(8) translation of polysemic words via the 'contextological dictionary', on basis of morphological and syntactic information and of semantic classes of SL words

(9) translation of any residual polysemes by 'best-match' equivalence table

(10) generation of TL morpho-syntactic information (case, number, congruence, etc.) on basis of TL grammatical information introduced at stage 6; sometimes involving further analysis of SL text (e.g. search for relative pronouns and identification of their gender, number and case congruence)

(11) morphological synthesis of TL word forms

It is evident that syntactic analysis is limited to the identification of grammatical categories (i.e. traditional parts of speech) necessary for Russian synthesis; there appears to be some check for syntactic compatibility but there is clearly no phrase structure analysis. There is also no distinction between levels of analysis: stages use mixtures of morphological, syntactic and semantic information as appropriate. Routines for analysis and synthesis are introduced whenever it seems best, e.g. synthesis of negation occurs in stage 6 before lexical translation (stages 7 to 9) and before relative clauses have been analysed in stage 10. In particular, the system is basically dictionary driven; the system uses a number of dictionaries (totalling 25,000 English words and 35,000 Russian entries): a 'common vocabulary dictionary', a 'general technical' dictionary, a specialised dictionary for 'computer science and programming' (the field in which AMPAR is operational), a dictionary of idioms, also divided by subject field, and a dictionary for Russian synthesis; but the

crux of the system is the 'contextological dictionary' for dealing with English words which have more than one possible Russian equivalent.

The English-Russian contextological dictionary was compiled on the basis of distributional and statistical analyses of English newspaper texts on political, economic and scientific topics amounting to some 700,000 words. The dictionary, compiled by students at the Moscow State Pedagogical Institute for Foreign Languages, was created to support English language teaching as well as the MT system (Marchuk 1979). Dictionary entries consist of algorithms which search to the left and right for particular words or grammatical categories or semantic classes. On the basis of such contexts the Russian output is determined. For example, the algorithm for English *issue*:

1	lft	1	list 1	3
2	tr		*nomer*	
3	rgt	1	periodical	5
4	tr		*nomer*	
5	rgt	1	list 2	7
6	tr		*vypusk*	
7	rgt	1	order (n)	9
8	tr		*otdacha*	
9	rgt	1	statement	11
10	tr		*opublikovanie*	
11	tr		*vopros*	

where the numbers on the right refer to the rule to be followed next if the condition does not apply; and where 'list 1' includes *today* and *yesterday*, and 'list 2' *note (n), currency*, and *bond*. As in the remarkably similar algorithms of Panov for the first Russian MT experiment (Ch.6.1)[4], output can sometimes be empty (e.g. for English prepositions). Although it is admitted that verbs and prepositions cause difficulties it is claimed (Marchuk et al. 1975) that cases of multiple meaning that can only be resolved by contextual analysis and require extra-linguistic information are relatively rare. Nevertheless, a statistical analysis of errors revealed that 10% of homographs were incorrectly resolved, mainly of the 'verb-noun' type (Kiselev 1984). The MT philosophy rejects syntactic analysis in favour of lexical context: "As is widely known, automatic recognition of syntactic structures is a task more complicated than resolving lexical ambiguity. Besides, the syntax depends largely upon the semantics of words in combinations and as such upon the lexical level" (Marchuk 1979) While many researchers may agree that semantics should play a much greater role in MT analysis than in some previous systems, not all would advocate a return to the statistical and contextual approaches of the earliest MT research.

Output from AMPAR is post-edited although the quality is considered good enough for unedited texts to be understood by specialists for "preliminary pilot information." There is work both on qualitative improvements of the system "based on a wider use of semantic and word combination properties", and on extension to other subject fields. It is claimed that for any new sublanguage "it is sufficient to supplement the system dictionary with 4-5 thousand Russian and English lexical units, and the word combination dictionary with 5-6 thousand dictionary entries" and that this can be completed by eight researchers in 3-4 months (Marchuk 1984)

18. 5: The NERPA system

The German-Russian system NERPA is based on the same linguistic and programming principles as AMPAR. The only major difference is the considerable role played by morphological analysis: German compounds not located in the dictionary are segmented and each part translated and synthesised into Russian phrases, e.g. *Informationsverarbeitung* becomes the noun phrase

[4] For an explanation see footnote 3 above.

obrabotka informatsii. NERPA came into experimental operation in 1981 covering the fields of computer science and programming (Marchuk 1984)

Both AMPAR and NERPA are designed as practical operational systems ('industrial' is the term used by the Russians), translating within relatively limited subject fields. Tikhomirov (1984) describes them as "multifunctional" systems, i.e. designed to translate polythematic documents, to be adjustable to any form of input, to provide interactive editing, to facilitate lexicographic research, and to permit prompt correction and upgrading. A unified software for AMPAR and NERPA has been designed to facilitate modification of programs by linguists; a 'special process control language' permits the easy insertion, correction, deletion and reordering of program modules, the alteration of dictionary information, the maintenance of files and the monitoring of the system. There are two translation modes running concurrently: the operational mode, and an upgrading mode for revision of the system and dictionaries and for evaluation of the effects of any changes before implementation in the operational system. (Marchuk 1984, Tikhomirov 1984)

18. 6: The FRAP system

The other MT project at the Centre for Translation is the French-Russian system FRAP, designed on different principles. This is apparently an operational development of the FR-II system of Kulagina at the Institute of Applied Mathematics (Kulagina 1976) The system is based on the 'transfer' approach, with the following familiar stages: morphological analysis, identification of verb phrases, dependency structure syntactic analysis, filtering out of semantically anomalous dependency analyses (some filters applied iteratively), syntactic transfer of all acceptable SL trees, substitution of SL lexical items by TL ones, syntactic synthesis and morphological synthesis. In essence, the system as described by Kulagina (1976) employed the 'filter' analysis approach seen in CETA (Ch.10.1 above) and the 'interlingual' LRC system (Ch.10.3)

Evidently, this basic design has been retained in FRAP. Marchuk (1984) describes it as having a "modular structure" divided into dictionaries and algorithms, and operating in several modes. The first is an "auxiliary word-for-word translation" enabling checks of the SL and TL dictionaries. The second mode, the 'syntactical mode' translates via syntactic representations and refers also to "the semantic component to verify the meaning of links and translated equivalents". The third mode is described as the "textual-and-semantic" mode, translating via a semantic representation "which may be accompanied by conciseness and semantic editing of the text content". The fourth mode, an "informational" mode, produces translated abstracts for a system of selective dissemination of information. Evidently FRAP is conceived, like AMPAR and NERPA, as a multifunctional system, integrating MT, lexicographical work and an abstracting service. The interface between syntactic and semantic representations has been designed but apparently not yet implemented. The nature of the semantic representation is not clear, but the main function of semantic analysis is to filter out "doubtful links found in the syntactical representation" (Marchuk 1984) The basic configuration of FRAP would appear to be on the standard 'transfer' model (Ch.3.9) and most similar to the TAUM approach, although details of the current system are meagre.[5]

The recent version of FRAP has been developed for three subject areas: electronics, computer science, and aviation and aircraft construction (Marchuk 1984). There are no indications of the quality of its output as yet. The earlier version FR-II (Ch.11.5) had been designed for translating French mathematics texts and had produced output which was evaluated on the same basis of CETA output (Vauquois 1975); a comparable quality of translation on a similar quantity of text was achieved: comprehensibility was judged to be "very good" by 61-63%, grammatical correctness "very good" by 49-54% and adequacy "very good" by 60-68% (Kulagina 1976).

---

[5] For a collection of articles on FRAP see: N.N.Leont'eva, et al. *Mashinnyi perevod i prikladnaya lingvistika: problemy sozdaniya sistemy avtomaticheskogo pervoda* (Moskva: Ped. Inst. Inostrannykh Yazykov, 1987).

18. 7: Other Soviet systems

It is clear that FRAP is now the most advanced MT system under development in the Soviet Union. Since 1976 the emphasis has been on practical systems of admitted low quality output. An example given by Marchuk (1984) is the system developed at the Chimkent Pedagogical Institute for the Kazakh Academy of Sciences which translates "British and American texts on chemistry and polymers". It is evidently no more than a 'dictionary translation' word-for-word system, but claiming to have "completely satisfied customers' needs for several years now".

There is evidence that Russians are beginning to exploit the potential of the computer for machine-aided translation. Marchuk (1984) mentions the work at the Minsk Institute of Foreign Languages on a databank of Russian-English phrases, "frequently encountered colloquial clichés used in stereotypical city conversational situations", with the intention of developing a microcomputer based translation system. As in Western Europe, there is considerable activity on large-scale term banks. The Centre for Translation in Moscow is creating term banks for English, French, German and Hungarian in the fields of computer science and aviation. It is also collaborating with Moscow State University on the MULTILEX computer dictionary (Russian, English, German, French), which provides on-line interrogation and output on screen or hard copy (Oubine & Tikhomirov 1982); in content and structure it is rather similar to EURODICAUTOM.

Interest in the 'sublanguage' concept derives naturally from the strong lexicographic tendencies of Soviet linguistic research. Marchuk (1984) traces the Soviet interest in the question to Andreev, and mentions a recent (1983) book by L.L. Nelyubin[6] on the topic which describes sublanguages in terms of four 'models' (functional-communicative, statistical, informational and linguo-statistical) and an English-Russian system, presumably experimental, based on these models for translating "organizational and management documents". Another example of the work being done is the research of Zubov (1984) at the Minsk State Pedagogical Institute of Foreign Languages, proposing to base MT systems on statistically derived characterisations of sublanguages.

While it is quite clear that theoretical studies are not neglected – there is evidently still interest in the 'combinatorial' dictionary (Ch.10.2), Marchuk (1984) – it would seem that the current emphasis in Soviet MT work is on practical bilingual systems for relatively restricted domains. In this respect it is in accord with the views of many Western groups, as Marchuk (1984a) is aware when he admits that "MT in the USSR is perhaps developing rather slowly but in an undeviating manner because other ways to overcome increasing language barriers in the USSR, as well as all over the world, are available".

18. 8: Research in Japan (1975-

After a period during the late 1960s and early 1970s when little original MT research was done in Japan, there has been a considerable revival in recent years. Some of this revitalisation can be attributed to the ambitious Fifth Generation Computer project of the Japanese government, which has included MT as one of its main objectives.[7] However, probably more important has been the Japanese 'language barrier': Japanese is an isolated language with no similarities to any other language, and despite popular impressions, the Japanese learn other languages with great difficulty; the need for good translation services is crucial to Japan's commercial and economic growth. The greatest demand is for translation to and from English, but French, Spanish, German and Chinese are also important. The Japanese realise that MT is only a partial answer; as in Europe, more immediately attractive are high quality text processing facilities and terminology databanks. In the latter respect, Japanese researchers admit that they lag a long way behind Western Europe and

[6] L.L. Nelyubin, *Perevod i prikladnaya lingvistika* (Moskva: Vysshaya Shkola, 1983)
[7] See below, Ch.19.3.

Canada, and the Kyoto University MT project (below) includes the creation of a large term bank for science and technology eventually to reach one million entries (Nagao 1984).

There has been and continues to be considerable research activity in Japan in the fields of computational linguistics and natural language processing. Much of the research has taken place within industrial and commercial companies or in joint university and government/industry projects and for this reason it has not been widely disseminated. In addition, of course, most of the documentation is available only in Japanese and for this reason inaccessible to most Western researchers.[8] Some idea of the extent of Japanese involvement in research on natural language processing, text analysis, linguistic databanks, and Japanese character processors can be found in the survey by Nagao and Tsujii (1980a). While the West may know little of Japanese research, the Japanese have followed developments in Western linguistics and computer science closely. An article written in 1979 by Tsujii (1980) reveals detailed familiarity with current research on artificial intelligence and linguistics in the United States and Europe. Nearly all of the most recent MT projects aim to explore various AI techniques, as we shall see, and many of the Japanese MT projects are part of larger research efforts within the fields of artificial intelligence, robotics and control systems; the Hitachi group is typical (Ishihara et al. 1985)

One of the major influences in MT system design has been the case grammar model of Fillmore (Ch.9.16), primarily because Japanese, as a language with 'surface' case indicators, relatively free word order and verb-final sentence structure, lends itself more readily to the fluid parsing of the case grammar approach than to the more rigid 'static' parsing of the phrase structure approach. However, the Japanese have not, on the whole, assumed that 'deep' cases (i.e. roles such as 'agent', 'instrument', 'location') are interlingual universals; indeed Tsujii of Kyoto University, who spent some time at GETA (Ch.13.3), has argued strongly against such an assumption. The general view seems to be that in view of the substantial syntactic and semantic differences between Japanese and, for example, English the most practical MT strategy is the 'transfer' approach.

18. 9: University of Osaka Prefecture

An example of a 'transfer' system using the case frame approach is the MT research project at the University of Osaka Prefecture (Nishida and Takamatsu 1982). Experimental systems have been developed for both English-Japanese and Japanese-English translation. In both, SL and TL 'transfer' representations are basically (semi-structured) lists of predicates and arguments with semantic case markers. For example, the Japanese sentence (with English lexical items in place of Japanese characters):

cars -(no) number -(ga) rapidly increase

(i.e. in English: *The number of cars is rapidly increasing*)

The SL interface representation is:

(PRED-ATTR.TRANS: increase, TENSE: present, ASPECT: progressive, OBJ-NUMber: number (NUM: *, OBJ-PHYS.OBJ: cars), MANNer: rapidly)

The Osaka system produces multiple parsings of SL input, checking all potential case frames for verbs in parallel. (The verb patterns were based on those given in Hornby's *Advanced Learner's Dictionary* (1963).) Rules for SL-TL structural transfer specify equivalent SL and TL sentence patterns, e.g. between a Japanese 'be' possessive and an English 'have' possessive:

(PRED-EXIST: is, OBJ-HUM: a daughter, LOC-HUM: with him)

i.e. *he -(ni wa) a -(no) daughter -(ga) is*

[8] Shortly after completion of this book, much more information about Japanese MT research in the 1980s became available. It is summarised in: W.J.Hutchins, 'Recent developments in machine translation: a review of the last five years', *New directions in machine translation. Conference proceedings, Budapest 18-19 August 1988*, ed. D. Maxwell, K. Schubert, T. Witkam (Dordrecht: Foris, 1988), pp.7-63. See also: M. Nagao, *Machine translation: how far can it go?* (Oxford: Oxford University Press, 1989), and 'Current machine translation systems in Japan', *Machine translation summit*, ed. M.Nagao (Tokyo: Ohmsha, 1989), pp. 213-224.

becomes:

(PRED-POSS: have, POSSESSOR-HUM: he, OBJ-HUM: a daughter)
i.e. *he has a daughter*

The English-Japanese system is claimed to have given "good results in the analysis of wide variety of English sentences." The Japanese-English system has given output such as the following:

Comparing two signals gives a measure of force exerted by fluid, and then an electronic circuit converts this measured value to a scale of a flow rate.

18.10 Hitachi

The research on the English-Japanese system at the Systems Development Laboratory of Hitachi exemplifies an empirical approach (Nitta et al. 1982, 1984). The Hitachi group argue against the rigid, detailed parsing of other MT systems (claiming that after all, few practical MT systems have resulted) and prefer looser heuristic parsing, akin to that of language learners who cope with "elementary grammatical knowledge". Their Heuristic Parsing Model is based on a "non-standard" grammar.

English sentences are segmented into Phrasal elements and Clausal elements (on the basis of punctuation, prepositions and conjunctions, rather like Wilks' method, Ch.15.1 above); then restructured as trees in which inclusive relations are distinguished from modifying relations (e.g. nouns in apposition, and noun phrase plus dependent prepositional phrase). The resultant structure is permuted, basically by pattern matching, and the appropriate Japanese equivalents are inserted with any necessary case suffixes and postpositions. The ATHENE system (Automatic Translation of Hitachi from English into Nihongo with Editing Support) is a deliberate attempt to explore low-level techniques of syntax-directed parsing, keeping to 'surface structures' and simple pattern matching and using a minimum of semantic information (mainly case frame valencies) for disambiguating dependency relations (Nitta et al. 1984). A large burden is thus placed on dictionary information, and the project has taken care to design lexicon structures and procedures which can be easily understood and applied by linguists unfamiliar with computer operations (Okajima et al. 1983). The primary emphasis of the ATHENE project has been on the syntactic analysis of English; the Japanese output is stylistically awkward word-for-word translation, although some refinement has been introduced by heuristic rules for the selection of Japanese postpositions. The limitations of the present system are indicated by the fact that problems of multiple meaning can be handled only during post-editing, and that the parser cannot deal with a number of ambiguous English constructions, such as the *ing*-form plus noun (either adjective plus noun or gerundive plus object) and the scope of coordination. Editing of output is clearly essential. A dictionary of some 70,000 entries has been compiled and the system has been tested on American economics articles from the *Wall Street Journal* and *Business Week*.

The group is exploring also Japanese-English translation. In this case a semantics-based approach has been adopted for analysing Japanese, essentially a case-frame parser producing a 'conceptual dependency' representation (Nitta et al. 1984, Ishihara et al. 1985). The Hitachi approach is characterised by practical realism: while syntax-directed parser may be best for English, semantics-based approaches are better for Japanese. Whereas the English-Japanese system was essentially based on the 'direct' syntactic transfer strategy (Ch.3.9), the Japanese-English system adopts an orthodox 'transfer' strategy. Both systems are written in PL/1.

18. 11: Fujitsu

Research on the ATLAS projects at Fujitsu combines work on a practical MT system using well-tried techniques and long-term work on a system using advanced AI methods (Sawai et al.1982). The advanced ATLAS/U system is designed as a high-quality multilingual transfer MT system applying knowledge representations and inference mechanisms. There are plans to

collaborate with the University of Stuttgart on a Japanese-German model (Ammon & Wessoly 1984)

The ATLAS/I is designed as a practical system translating a limited range of materials from Japanese into English. Essentially it is still experimental, but an operational version has been handling translations of software reports for the company. The translation process goes through the following stages. 'Preprocessing' segments the Japanese text on the basis of punctuation mark; then follows consultation of the SL-TL dictionaries: 'Phrase dictionary reference' (for idioms and compounds), 'Word dictionary reference', 'Adjunct analysis' (which identifies Japanese postpositions and assigns codes), and 'Proper noun processing' (which gives the semantic marker for names to any word not found in the dictionary). Then follows 'Syntactic analysis' (identifying noun phrases and their relations), 'Surface case analysis' (involving nouns and their postpositions), 'Deep case analysis' (on the basis of patterns of verb case frames and 'surface' cases), 'Structural translation into English' (via a Japanese-English pattern translation table), 'Synthesis of English sentence' (invoked by the selected pattern, and including the insertion of prepositions, articles and plural endings), and finally 'Morphological synthesis'.

A distinctive feature of the system is the use of a uniform representation of case frames, syntactic conditions and English synthesis rules. This 'frame knowledge representation' (based on AI ideas, cf. the similar concept in METAL, Ch.13.4) facilitates the expandability of the system. The number of sentence patterns it can handle in the structural transfer program is being gradually increased by additional rules. In 1982 the dictionary of the operational system (translating software reports) contained 5000 words and there were 400 grammar rules. The system evidently depended to some extent on interactive assistance at various stages of analysis in order to handle structures and vocabulary not yet included.

18. 12: Nippon Telegraph and Telephone Corporation

More ambitious in conception is the project at the Musashino Electrical Communication Laboratory of the Nippon Telegraph and Telephone Corporation (Nomura 1983, Iida et al. 1984, Naito et al. 1985) The LUTE project (Language Understander, Translator & Editor) is a good example of an AI-based MT experimental system. The goal is high-quality Japanese-English and English-Japanese translation. The prototype LUTE system was developed in May 1982. Basically, it is a bi-directional 'transfer' system for English and Japanese. The linguistic core of the system is an extensive case frame analysis of Japanese and English, and of contextual "world knowledge". Text processing of SL input involves the matching of case frames and knowledge representations against lexical information. The result is a structure of semantic and syntactic dependencies. Transfer from SL to TL representations is by pattern matching and frame manipulation. The final generation of TL texts is envisaged as simple linearisation of TL frame structures. The distinctive features of LUTE are the extension of case frame structures to tense, aspect, modality and semantic implicatures, in part adopting Montague grammar approaches (Naito et al. 1985), and the intended integration of 'common-sense' and 'expert' knowledge databases for SL text disambiguation. The system is implemented within a highly sophisticated environment for interactive text processing, screen editing, and integrated dictionary and grammar development. As in other Japanese projects, programs are to be written in LISP (the Maclisp dialect in this case.) The system is being developed initially on a corpus of general scientific articles from *Scientific American* and its Japanese equivalent. There is evidence of considerable ingenuity and expertise in both linguistic and computational aspects of the experiment; LUTE is an example of Japanese skill in integrating the most advanced techniques in a project at the forefront of current research.

18. 13: Kyoto University

There had apparently been some activity at Kyoto in the first decade of MT research (Ch.7.1), but effectively the major effort began around 1968 with the experimental project under

Toshiyuki Sakai. The project developed two small-scale 'syntactic transfer' systems for English-Japanese and Japanese-English translation. Syntactic analysis was on the phrase-structure model, transfer was purely via syntax pattern matching, and there was no semantics. "Distinction between singular and plural form is not considered, and articles are neglected" The rudimentary results are well illustrated by two example translations (Sakai et al. 1970): *Passing through long tunnel of border, it was snow country; We cross river that water flow.*

At the present time there are two major MT projects at Kyoto University. The first is the development of a practical operational system for the translation of English article titles into Japanese. The second is a full-scale government-sponsored project for English-Japanese and Japanese-English translation, initially of abstracts. In addition, there have been some small-scale experimental projects in recent years.

The TITRAN system has been designed specifically for the translation of titles of scientific and engineering papers from English into Japanese (Nagao and Tsujii 1980, Nagao et al. 1982) The system relies heavily on the restricted grammar and vocabulary of titles. It was found that more than 98% of titles consist of noun phrases only. Therefore, the system has been designed to translate only noun phrases and prepositional phrases; it can deal with infinitive verbs, verbal adjectives ending in *-ing* or *-ed* but it cannot treat embedded relative clauses containing finite verb forms. The semantics of titles are equally restricted. It was found that nouns are usually specific terminological words of a particular field, and so there is little problem with polysemy. Compounds could be treated without problems as units, and it was found that simple noun phrases could be translated word for word without needing any change in word order. There are six stages of translation as follows:

(1) Dictionary lookup and search for idioms
(2) Parsing of conjunctive phrases
(3) Parsing of simple noun phrases by a transition network parser
(4) Handling of special structures and semantic disambiguation
(5) Skeleton pattern matching and word order change for Japanese
(6) Synthesis of Japanese title

Semantic disambiguation is limited to recognition of basic case frame relations, i.e. the compatibility of verb participles with particular classes of nouns (only five were found necessary: tool, aspect, physical object, theory, unit.) Hence, the system can distinguish between *-ing* forms which may be modifiers of 'tool' nouns (*measuring devices*) and *-ing* forms which govern object nouns (*measuring temperature*). For structural transfer it was found that the variety of patterns was remarkably limited; only 18 different ones have been identified and the vast majority of titles conform to just three of these patterns (N, N+prep+N, N+prep+N+prep+N). However, the Skeleton pattern step does not cover all transfer operations; certain structures have to be treated by special routines (in step 4), such as titles which are prepositional phrases, e.g. 'On pattern recognition'. In these cases, the noun phrase is translated independently and the preposition is transposed into a Japanese postposition at the end of the phrase. Finally, it is admitted that certain structures cannot be handled satisfactorily, in particular coordination.

Within its limitations the system has had a satisfying success rate. In the specific area of physics and mathematics from the INSPEC database the average success rate was as high as 93%; out of 3000 titles there were only 42 rejected (1.4%), primarily because they included finite verb forms, and only 5% were either wrongly translated or could not be understood. Even when extended to the full range of scientific and engineering papers covered in the INSPEC database the average was 80%, and this was attributable primarily to inevitable deficiencies in the dictionaries.

The system is now being used on a trial operational basis at the Tsukuba Research Information Processing System (RIPS) of the Agency of Industrial Science and Technology. In a collaborative project with Saarbrücken, the system is being linked to SUSY for an experiment in the translation of titles from and into German (cf. Ch.13.2, Ammon and Wessoly 1984).

At the same time as Nagao and his group were developing TITRAN, various small-scale MT projects were active. A Japanese-English system for translating computer software manuals investigated the use of case frames for Japanese verbs in a 'lexicon-driven' approach to analysis. There was substantial work on the problems of morphological analysis and dictionary lookup, with of course particular attention to the problems of Japanese character handling in input and output (Nagao & Tsujii 1980) More speculative and tentative was the research by Nishida & Doshita (1982) on an experimental interactive English-Japanese system which investigated the use of a logico-semantic 'interlingua' on the basis of Montague's semantic theory (Montague 1974), and which has been continued by Tomita at Carnegie-Mellon University (Ch. 17.13) The research was part of the general interest at Kyoto on the possibilities of incorporating 'understanding' routines in MT systems; the prototype, written in LISP, was tested on four technical reports and computer manuals. Equally speculative are Nagao's suggestions for a 'learning' MT system: on the basis of example sentences presented to it, the system would develop its own methods of analysis. It would work by "example-guided inference, or machine translation by analogy principle".[9] It is argued that use of the computer for gathering and analysing linguistic data is a more fruitful approach than application of linguistic models and methods: "Linguistic theories change rapidly to and fro, and sometimes a model must be thrown away in a few years. We will rely on the primary data rather than analysed data which may change sometimes because of changes in the theory" This statement, though it recalls the views of earlier MT 'empiricists', may in fact have more to do with the Western bias of theoretical linguistics and the 'failure' to produce models suitable for non-Western languages, cf. Ch.19.2 below. As just one instance, there is the view of another Kyoto researcher, Tsujii, who from his experience at GETA (Tsujii 1982) while working on a trial English-Japanese system, became convinced that there was no real justification for assuming a 'universal' set of case relations, that it was necessary to establish different types of classification for English and Japanese verbs, and that, therefore, transfer had to be lexicon-driven.

The national MT project of the Japanese government, the Mu-project, began in April 1982 (Nagao 1984, Nakamura et al. 1984, Nagao 1985, Nagao et al. 1985).[10] The project, called officially "Fast Information Transfer System for Japanese & English Documents in Science and Technology", includes the creation of a Japanese-English terminology databank for science and technology (beginning with electrical engineering), the development of a 'transfer' MT system for Japanese-English and English-Japanese translation of scientific and technical abstracts and an integrated system incorporating both databank and MT system. The Japan Information Center of Science and Technology (JICST) has responsibility for the development of the term bank; the Tsukuba Information Center (RIPS) for the integrated system and for overall evaluation; and the Electrotechnical Laboratory (ETL) in collaboration with Kyoto University for the MT software, the grammars and the linguistic information for dictionaries.

The method will follow the familiar 'transfer' model (Nagao 1984): series of subgrammars, partial analyses, and 'deep structure' interface representations based on case grammar analyses (as we have seen, the Japanese are in general agreement that the 'case grammar' model is the most suitable for the Japanese language) There will be no attempt to strive for interlinguality in transfer representations and procedures; much of the transfer will be controlled by specific instructions contained in SL and TL dictionary entries ('dictionary rules'), which will take precedence over any general transfer rules, i.e. it is basically a "lexicon-driven machine translation system". The detailed

---

[9] This proposal for what has subsequently been known as Example-Based Machine Translation was made originally in 1981 but not published until 1984; see: M.Nagao 'A framework for a mechanical translation between Japanese and English by analogy principle', *Artificial and human intelligence: edited review papers presented at the international NATO symposium on Artificial and Human Intelligence... Lyon, France, October 1981*, ed. A.Elithorn and R.Banerji (Amsterdam: North-Holland, 1984), pp. 173-180.

[10] See also: M.Nagao, J.I.Tsujii, and J.I.Nakamura, 'Science and Technology Agency's Mu machine translation project', *Future Generations Computer Systems* 2, 1986, pp. 125-139.

semantic analyses of Japanese and English verbs and their case frames is the responsibility of the Kyoto team; for the scientific and technical vocabulary the JICST term bank is expected to provide appropriate information. It is intended to classify vocabulary by subject fields to assist with polysemy, and it is suggested that a thesaurus might be added in the future.

Although translation in both directions is being planned from the start there will be in fact two independent systems: one for Japanese-English, and one for English-Japanese. "To achieve high-quality translation, there is the possibility that special and different information and mechanisms are required for each translation direction. Japanese analysis grammar, for example, may not necessarily be the same as Japanese generation grammar"; it is thought possible that "one common set of transfer rules for both directions may be feasible. However, to play it safe, we are generating two independent sets of transfer rules at present, because we have still little confidence in this reversibility of the rules." (Nagao 1984) In this respect the project team has learnt from the experience of earlier 'transfer' systems, e.g. GETA and SUSY (cf. Ch.13.2-3).

The software is to be written in LISP, with which the Kyoto team has 20 years of experience. It includes GRADE, a metalanguage for specifying grammar and dictionary rules and for interactive MT developmental research (Nakamura et al. 1984, Nagao et al. 1985). GRADE is a powerful grammar writing system (GRammar DEscriber) written in UTILISP (University of Tokyo Interactive LISP), providing a framework for writing programs of analysis, transfer and synthesis, handling word-specific grammar rules, which are able to treat syntactic and semantic ambiguities, and for writing transduction rules for labelled trees. The conception is very similar to the developments in GETA and Eurotra: tree-tree conversions, specifying conditions, features, partial analyses, and output representations; networks of subgrammars, flexible control, etc.

The goal of the project is to achieve with four years a practical working system for "the batch translation of abstracts at JICST, etc. for collective post-editing, and an on-line conversational mode for writing research papers in English via machine translation".[11] It is assumed that post-editing will be essential for the foreseeable future, and the total system configuration remains uncertain, partly because until now the Japanese have had no practical experience at all of operational MT systems.

18. 14: Other Japanese projects

There are known to have been many other MT projects in Japan. For example there was at Kyushu University a small-scale 'transfer' Japanese-English system for translating texts on transistors which included analysis of case relations (Shudo 1974). Elsewhere it is known that research continues at Kyushu University (based on dependency grammar) and many new projects have been established in recent years in Japan: at the Tokyo Institute of Technology under H.Tanaka, at the Electro-Technical Laboratory under S.Ishizaki (a system based on Schank's MOPTRANS theory, cf. Ch.15.2), at the National Electric Company, at Toyohashi University of Technology, and at the computer manufacturers Toshiba, Oki, Mitsubishi, and IBM-Japan (Nishida 1985). No doubt recent interest in the possibilities of AI will encourage more projects to be set up in the near future.

[11] For a description of the JICST system when installed see: T.Ashizaki, 'Outline of the JICST machine translation system', *MT Summit II, August 16-18, 1989, Munich* (Frankfurt a.M.: Deutsche Gesellschaft für Dokumentation, 1989), pp. 44-49.