Chapter 12: Direct translation systems since 1965

12.1. Systran system of Latsec Inc./World Translation Center (1968 -)

Although involved in the installation of the Georgetown systems at the Oak Ridge National Laboratory and at Euratom in Ispra, Italy (Ch.4.3), Peter Toma had already set up a company to pursue MT research. This was Computer Concepts Inc. based in Los Angeles. By April 1963 he reported the development and operational implementation of AUTOTRAN a "fast efficient and accurate" MT system for Russian-English translation with a dictionary of 100,000 stem entries in the fields of atomic energy and medicine, programmed on the IBM 7090 (*CRDSD* 10-11, 1962; 13, Nov 1964).

Shortly afterwards in 1964 Toma moved to Germany, where he was able to continue his research with the support of the Deutsche Forschungsgemeinschaft (DFG). Here, on the IBM 360/50 at Bonn University, Toma was able to begin the development of the Systran Russian-English system (again working mainly at nights, according to his own account, Toma 1984). A detailed description of this prototype, showing clear signs of its GAT-SERNA ancestry, has been given by Janda et al. (1970). In 1967 the DFG had put forward a proposal to develop a Russian-German version at the University of the Saarland; the proto-Systran Russian-English system was thoroughly evaluated, but eventually the Saarbrücken group decided to develop its own system (Ch.13.2 below)

In 1968 Toma founded his own company Latsec Inc. in La Jolla, California, in order to continue development (now also supported by the US Air Force) and by early 1969 Systran was ready for testing at the Wright-Patterson Air Force Base (Dayton, Ohio). From July 1970 Systran took over from the IBM Mark II system (Ch.4.2), where it continues to provide Russian-English translations for the USAF's Foreign Technology Division to this day. Subsequently, Systran was used by the National Aeronautic and Space Administration (NASA) during the joint US-USSR Apollo-Soyuz space project (1974-75), and in 1976 it replaced the Georgetown system at Euratom.

For NASA Latsec Inc. developed an English-Russian version, apparently within one year (Bruderer 1978). At the same time, other language pairs were being prepared. A Chinese-English version was demonstrated in 1975 to representatives of the American government, and a German-English version to the US Army.

The most significant development, however, was to prove to be the English-French system (started in 1973). This version was demonstrated in Luxembourg in June 1975 to representatives of the Commission of the European Communities (Bruderer 1978). As a result, the Commission concluded a contract with Latsec Inc., now renamed the World Translation Center, to develop versions for translation between languages of the European Communities. The English-French version was delivered in February 1976, followed by the French-English version in 1978 and the English-Italian version in 1979. All three systems have been under continuous development and expansion by staff of the European Communities at Luxembourg, and all came into full production in March 1981 translating internal documents of the European Communities on an ever increasing scale (see ch.14.1 below)

There are now a number of other users of these versions: e.g. the French-English system at the Centre for Nuclear Research (Kernforschungszentrum) in Karlsruhe (Habermann 1984) and Aérospatiale in Paris; the English-French system at General Motors of Canada (Sereda 1982), Xerox (Elliston 1979, cf.17.2 below) and Aérospatiale; and the English-Italian system at General Motors of Canada. There will certainly be others. Subsequently, Systran has also brought to operational status an English-Spanish system (purchased by General Motors of Canada and Xerox, among others) Other versions are said to be at an "advanced stage of development": English-German, German-English, English-Portuguese, English-Arabic, German-French, and German-Spanish; and plans are said to be under way for Spanish-English and Japanese-English (Van Slype 1983, Pigott 1984)

In 1978 the World Translation Company Canada was established to market Systran II (a system integrating MT programs, word processing and photocomposition systems) in the United States, Canada and parts of Europe. In 1979 the Systran Institut was set up in Munich to promote Systran in Europe, since when a number of contracts with companies have been concluded. Finally, a Systran service bureau has been established in Luxembourg (after an abortive attempt in 1980 to open one in Canada (Van Slype 1983) which can provide translations to companies which cannot purchase their own system. Having started as a project for US military purposes, Systran is now a commercial product.

General descriptions of Systran are to be found in Toma 1977, Whitelock & Kilby (1983) Van Slype & Pigott (1979), Pigott (1981), and briefer descriptions by Toma (1974, 1976a, 1976b). For the Russian-English system the basic material is in the technical reports (Toma et al. 1970, 1972, 1973, 1974).[1]

In many respects, Systran may be regarded as essentially a greatly improved descendant of the Georgetown 'direct translation' system. Linguistically there is little advance, but computationally the improvements are considerable, resulting in the main from the 'modularity' of its programming design. The monolithic complexity of the Georgetown system is overcome and the modification of any part of the processes can be undertaken with much greater facility and with much reduced risks of impairing the system's overall efficiency. Modularity is reflected in the following features. There are two main types of programs: (i) system programs, written in assembler code, which are independent of particular languages; these are control and utility programs, such as those responsible for dictionary lookup routines; and (ii) translation programs which are broken down into a number of stages, each with separate program modules. Translation programs for SL analysis and TL synthesis are to some extent independent of particular SL-TL pairs, although there are still a number of procedures during SL analysis which are determined by the needs of a particular TL. Nevertheless, the modularity of the translation programs has enabled the relatively straightforward introduction of new techniques of analysis wherever they seem appropriate.

However, the main component of the system remains the large bilingual dictionary containing not only SL-TL lexical equivalences but also grammatical and semantic information used during analysis and synthesis. Much of this information is formulated as algorithms to be invoked during various stages of the translation processing. As in 'direct' systems (like Systran's ancestor, the Georgetown system), these SL-TL dictionaries are of great complexity, and so consistency and regularity is difficult to maintain. While it is therefore true to characterise Systran as a partially 'transfer' system in that programs of structural analysis and synthesis are largely independent, the main translation processes are driven by the SL-TL dictionaries, as in 'direct' systems. Systran may therefore be regarded as a hybrid 'direct-transfer' system (Fig.21).

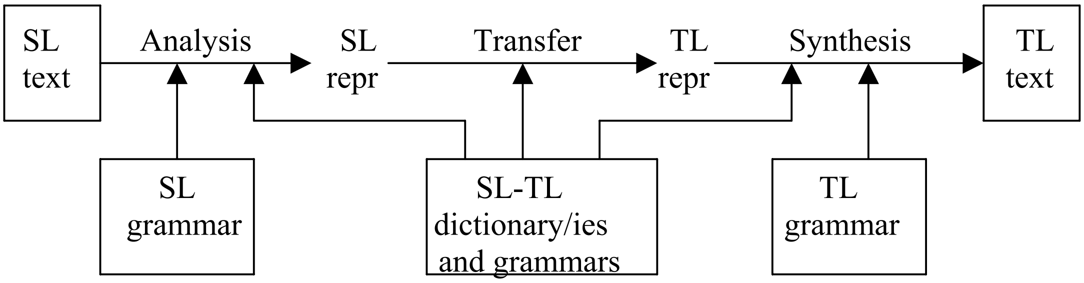

Fig. 21 – Hybrid 'direct-transfer' system

[1] For later developments see chapter 10 in Hutchins, W.J. and Somers, H.L. *An introduction to machine translation* (London: Academic Press, 1992)

Toma's (1977a) claims that Systran is a multilingual system are true, therefore, only with respect to the generality of the programming for structural analysis and synthesis. New versions of Systran are developed by incorporating SL and TL specific procedures in analysis and synthesis and by compiling new SL-TL bilingual dictionaries. Work on dictionaries does not, however, have to always start from scratch as in many cases there need be only minor differences in the coding of SL lexical items when coupled with a new TL (Pigott 1983)

The dictionary database for Systran comprises two bilingual dictionaries, one of single-word entries and one of multi-word expressions. From these are derived the dictionaries used during translation: a High Frequency dictionary including prepositions, conjunctions, irregular verb forms, etc. and the first words of idiomatic expressions; a Limited Semantics dictionary for idioms and compound nouns to be treated as lexical units; a Conditional Limited Semantics dictionary for dealing with semantic compatibilities and valencies; and the Main dictionary divided into stems and endings (except in the case of English). Dictionary entries include codes indicating: morphological class (verb and noun paradigms), part of speech, government and valency, agreement, transitivity, noun type (animate, mass, abstract, etc.), semantic markers ('physical property', 'food product', etc. - 450 such codes are available), any TL preposition governed by the item, and TL equivalents with morphological and syntactic information for synthesis. In Systran there are five basic stages (Whitelock & Kilby 1983, Van Slype & Pigott 1979): Input, Main dictionary lookup, Analysis, Transfer and Synthesis (fig.22)

The Input program loads the text (in Roman transliteration, in the case of Russian) and checks each word against the High Frequency dictionary. In the next stage, the remaining words of the text are sorted alphabetically and searched for in the Main Stem dictionary. The words are then sorted back into the original text sequence and passed to morphological analysis (in the case of languages like Russian and French) before going to the Analysis program proper. When English is SL, there is no morphological analysis as the dictionaries contain full forms; but for Russian and French as SL stems and endings are entered separately.

Syntactic analysis consists of seven 'passes' through the SL text:

(1) The resolution of homographs by examination of the grammatical categories of adjacent words (for English 83 different types of homograph have been identified).

(2) Identification of compound nouns (e.g. *blast furnace*) by checking the Limited Semantics dictionary (in some versions of Systran this pass precedes homograph resolution.)

(3) Identification of phrase groups by searching for punctuation marks, conjunctions, relative pronouns, etc. (i.e. a rudimentary phrase structure analysis)

(4) Recognition of primary syntactic relations such as adjective-noun congruence, noun-verb government and noun-noun apposition; this pass is performed in a right-to-left scan.

(5) Identification of coordinate structures within phrases, e.g. conjoined adjectives or nouns; this pass makes use of semantic markers to establish acceptable conjunctions, e.g. in *smog and pollution control*, the coordination of *smog* and *pollution* rather than *smog* and *control*.

(6) Identification of subjects and predicates, principally by searching first for a finite verb and then for a (preceding) noun not already marked as an 'object' or 'modifier'.

(7) Recognition of prepositional structures, searching first right-to-left for a preposition and then left-to-right for its dependent noun phrase.

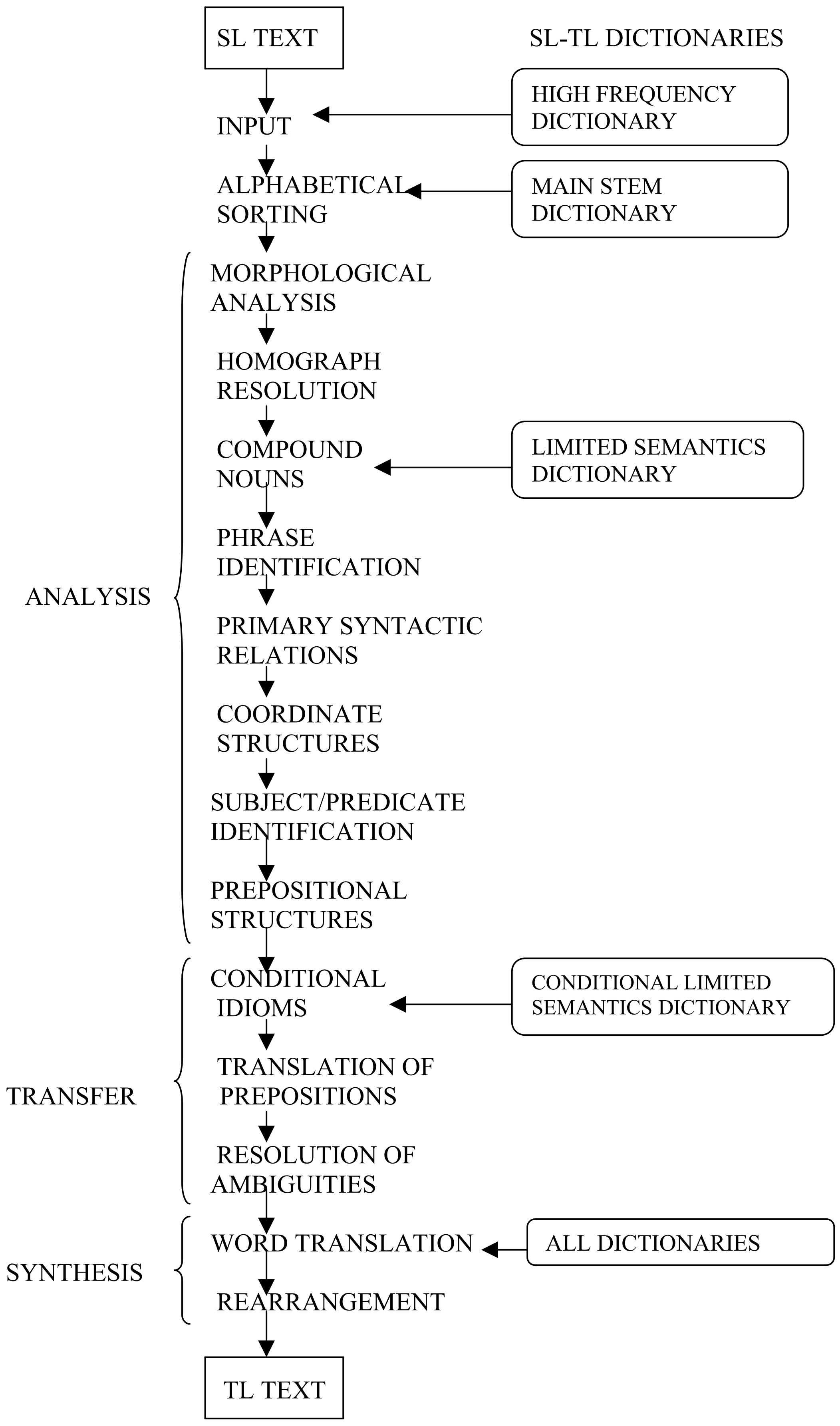

Fig.22 – Systran stages of translation

The Transfer program has three parts:

(1) Search for words with idiomatic translations under certain conditions, (e.g. if *agree* is in the passive, it is translated as French *convenir*, otherwise it appears as *être d'accord*, by reference to the Limited Semantics Dictionary.

(2) Translation of prepositions, using the semantic information assigned to words which govern them and which are governed by them.

(3) Resolution of the remaining ambiguities, generally by tests specified in the dictionaries for particular words or expressions.

The last stage, Synthesis, produces sentences in the target language from the equivalents indicated in the dictionaries, modifying verb forms and adjective endings as necessary, and finally rearranging the word order, e.g. changing an English adjective-noun sequence to a French noun-adjective sequence. An additional module in French synthesis is a routine for dealing with the selection of pronouns; the routine anticipates the possibility that the subject pronoun of a following sentence may refer anaphorically to the subject noun of the sentence it is dealing with. In this way English *it* may be correctly translated as *il* or *elle*.

The latter is one example of the improvements made possible by Systran's modular programming. Another example is to be found in the parsing program. As in other systems, most attention is paid to the identification of 'surface' phrase structure relations (e.g. noun phrases) and dependency relations (e.g. nouns and adjectives). For the identification of subjects and predicates (at pass no.6), however, somewhat 'deeper' analysis has been included. For example (Billmeier 1982), the passive sentence *The texts were translated by a computer* would be analysed roughly as:

```
Sentence
        Predicate: verb, past passive................................translate
             Deep subject:.............................computer
             Deep object:...............................texts
             Subject: noun................................................texts
                     Determiner: def.art..........................................the
             Prep. phrase 1: preposition............................by
                     Noun phrase: noun..........................................computer
                             Determiner: def.art............................................a
```

Likewise, a noun phrase containing 'deep' subject-predicate relationships would receive a parallel analysis. Thus, the phrase *the translation of texts by computer* would be analysed roughly as:

```
Sentence:
     (Subject): verbal noun..............................................translation
     Deep subject:......................................computer
     Deep object:........................................texts
     Determiner: def.art.....................................................the
     Prep. phrase 1: preposition.......................................of
             Noun phrase: noun..................................................texts
     Prep. phrase 2: preposition......................................by
             Noun phrase: noun..................................................computer
```

Consequently, Systran is producing analyses which mix 'surface' and 'deep' structure information (rather similar indeed to analyses in TAUM and GETA in this respect, see Ch.13.1 and

13.3 below); however, as Billmeier (1982) points out, Systran's incorporation of such features is selective, partial and not based on a coherent linguistic model.

The description of Analysis and Transfer as separate stages may be a little deceptive, in that it is not only during Transfer that TL selection takes place; in particular the second pass which treats 'idiomatic' compounds decides on TL forms, which in turn determine later stages of analysis. The distinction between analysis and transfer is less clear-cut in Systran than in 'purer' examples of 'transfer' systems, cf.3.9 above (Whitelock & Kilby 1983)

The main burden of the translation process rests, as pointed out earlier, with the large bilingual dictionaries. The information which is assigned to SL lexical data is restricted to that which is found necessary for facilitating conversion into one specific TL. Any information about either the SL or TL lexicon or grammar is included, in any mixture which seems convenient. As a consequence, there seems to be no uniformity, methods are inconsistent, coverage and quality are uneven, and modifications of one section of the dictionary can often have unexpected consequences.

Some examples of the mixture of SL and TL information can be found in the Russian-English system. The routine for inserting definite and indefinite articles combines syntactic information about the Russian text (e.g. whether the noun is qualified by a following genitive noun, a prepositional phrase or a relative clause), some semantic information (e.g. whether the Russian is an ordinal number) and information on English equivalents (e.g. an English 'mass' noun such as *water* usually requires a definite article). The methods of determining English syntactic forms are varied: in some cases, they are controlled by codes in Russian lexical items, e.g. *esli* includes a code to change a Russian infinitive construction ('if to examine...') to an English finite form (*if we examine...*) (Toma et al. 1970); in other cases the English syntactic form results from a manipulation of the output, e.g. Noun + *of* + Verbal noun + *of* + Noun (*result of treatment of burns*, from RESUL6TAT LECENI4 OJOGOV) becomes Noun + *of* + Gerundive + Noun (*result of treating burns*) (Toma et al. 1973).

The bilingual dictionaries make use of semantic categorisation, both as indicators of subject fields (to assist the resolution of polysemy) and as indicators of semantic compatibilities. However, the system of 'semantic classification' is so complex that at times it appears to be *ad hoc*. Examples of the use of 'semantic classes' from the Russian-English version are: e.g. the translation of Russian prepositions according to the 'semantic class' of adjacent verbs or nouns: *DO* is translated as *up to* if the preceding verb or noun is '+increase' and as *down to* if it is '+decrease', *PO* is translated as *along* if the following noun is '+linear', as *over* if '+nonlinear' and as *using* if '+metal tool' (Toma et al. 1974). Similarly, Russian 'noun + genitive noun' structures may be translated as English 'noun + noun' (where the first noun modifies the second and corresponds to the Russian genitive) only if the 'semantic class' of the Russian genitive is COMP(osition), MATH(ematics), MEAS(ure), MOTION, OPTICS, QUAL(ity), etc. (Toma et al. 1973). Clearly, these 'semantic classes' have nothing to do with the semantics of Russian, they are *ad hoc* labels (some indicating 'components of meaning', others subject fields) designed solely to overcome difficulties with the English output. Similar problems with Systran semantic classification was encountered in the development of the English-French system by the European Communities (see Ch.14.1)

The Russian-English version has been in regular use by the US Air Force since 1970. It is reported that at the present time, over 100,000 pages are being translated every year, more than 2 million words in 1980 (Van Slype 1983). Most texts are used unedited, mainly for 'information scanning', only 20% receiving 'light editing'. There is said to be "considerable user satisfaction" and a 90-95% accuracy level is claimed.

It is only recently, however, that the Russian-English version has been evaluated at all systematically. The 1972 report by Leavitt et al. (1972) was concerned mainly with cost analysis and the optimisation of post-editing and recomposition functions. Sinaiko (1971) compared a 1964 Mark II translation and a 1971 Systran translation of the same Russian text (itself a good translation of an English article), finding that Mark II left 1.2% of the words untranslated and Systran 2.3% and that Mark II provided alternative translations for 6.3% of the words and Systran for 5.3%. He concluded that “little progress has been made in recent years” on MT systems.

Later output from Systran has shown, however, that this conclusion was too harsh, as a comparison of the following (perhaps untypically good) Systran output with Mark II translations (Ch.4.2 above) will show (Bruderer 1978):

> THE SWISS PUBLIC IS WORRIED, THE BASEL NEWSPAPER “"NATIONALZEITUNG” WRITES IN ONE OF THE LAST ISSUES. RECENT AMERICAN STATEMENTS ABOUT THE FACT THAT THE USA CAN USE FORCE IN THE NEAR EAST, THE NEWSPAPER EMPHASIZES, CAUSE ALARM ALL OVER THE WORLD. AS CONCERNS SWITZERLAND, THEN, IF THIS COURSE CONTINUES, IT WILL EXAMINE THE QUESTION CONCERNING AN EXIT FROM THE RECENTLY CREATED ON THE INSISTENCE OF WASHINGTON INTERNATIONAL ENERGY AGENCY, WHICH UNITES A NUMBER OF THE CAPITALIST COUNTRIES – THE GREATEST USERS OF OIL.

A thorough evaluation of the Russian-English system by Knowles (1979) was based on translations produced during tests in 1977 conducted by the Gesellschaft für Mathematik und Datenverarbeitung in Bonn (1977). The corpus comprised a pedagogic grammar of Russian written for German students (2000 sentences), and four Russian technical texts on scales and weighing, airports, helicopters, and eyesight (500 sentences) taken from Great Soviet Encyclopaedia. Some of the translations from the grammar were undoubtedly odd: *We heard, as a door was discovered and as someone entered into an adjacent room* (i.e. ‘We heard a door opening and someone entering the next room’), *They speak, as if it left* (i.e. ‘They say he has gone’) – there were particular problems with pronouns because of the lack of context. Results from the technical texts were somewhat better:

> HELICOPTER, A FLIGHT VEHICLE HEAVIER THAN AIR WITH VERTICAL BY TAKEOFF AND LANDING, LIFT IN WHICH IS CREATED ONE OR BY SEVERAL (MORE FREQUENT THAN TWO) ROTORS... A HELICOPTER TAKES OFF UPWARD VERTICALLY WITHOUT A TAKEOFF AND IT ACCOMPLISHES VERTICAL FITTING WITHOUT A PATH, MOTIONLESSLY “WILL HANG” ABOVE ONE PLACE, ALLOWING ROTATION AROUND A VERTICAL AXIS TO ANY SIDE, FLIGHT IN ANY DIRECTION AT SPEEDS IS PRODUCED FROM ZERO TO THE MAXIMUM...

Although Knowles found that errors were occurring on average every four or five words, many of these could be easily rectified by additional entries in the dictionaries or by additional grammatical information (particularly on valency relationships). Knowles also suggested that further improvements might be possible with a Wilks-type ‘semantic calculus’ to identify anaphoric relations (Ch.15 below), and by more sophisticated ambiguity routines. In his view, greater consistency and perseverance would overcome many of the errors, since with Systran’s modularity and open-endedness it should be possible to incorporate enhancements without undue difficulty.

The USAF Russian-English system has been under constant development since its installation in 1970 (Bostad 1982) During a five-year period, the number of homograph routines doubled and stem entries in dictionaries increased by 30,000. Although there was evidence of improvements in the quality of translations, there was also a growing concern that “a certain degree of degradation was occurring”, after ‘improvements’ had been introduced, in other parts of the system errors were appearing where none had occurred before. Monitoring changes revealed that, on average, the "improvement/degradation ratio was consistently around 7:3". On aggregate there was progress, but it was not uniform and there were substantial losses. The answer was to take greater care; every proposal was checked against a library of Russian input texts (ca. 50 million words), and accepted only if there is no degradation. Improvement of the USAF system is therefore now a matter of "fine tuning" the quality, not of making large scale modifications.

In many respects, the Systran Russian-English system dates back to the early 1960s in conception. The later Systran versions for English-French, English-Italian, and other pairs incorporate improvements resulting from experience with Russian-English (Ch.14.1). Whether, as a consequence, these newer versions can avoid (or at least delay for a longer period) the quality degradation experienced by the US Air Force model has yet to be seen.

12.2. Logos Development Corporation (1969-78)

The Logos Development Corporation was founded in 1969 by Bernard E. Scott in order to continue research under the sponsorship of the US Air Force for an English-Vietnamese MT system. He had begun preparatory work on the system in the Spring of 1965 at Computer Technology Inc. In June 1970 a public demonstration of the system (LOGOS I) was given on a small corpus of just 1300 words, which was considered sufficiently satisfactory for the US Air Force to commence the translation of training manuals and to recommend further development. Nevertheless, the designers described it as an "initial capability system... not yet sufficiently developed for a large scale machine translation of technical material" and emphasised that LOGOS output required considerable post-editing. (Byrne et al. 1970)

LOGOS I maintained complete separation of SL analysis and TL synthesis and of computer programming processes and linguistic data and rules. It was claimed that "in theory" the system could be "applied to translation problems between most language pairs." This was because the "programs are entirely language-independent" even though the procedures themselves and the dictionaries were specifically designed for one particular SL-TL pair, English and Vietnamese. In brief, LOGOS I was a hybrid 'direct-transfer' system (Ffig.21), in which bilingual dictionaries are coupled with separate defined stages of analysis, transfer and synthesis.

No pre-editing of the English text was necessary. The program automatically detected word and sentence boundaries and identified parenthetical sequences. There were two dictionaries, one of high frequency words loaded into computer memory at start up, and a large dictionary on tape searched sequentially (eventually containing some 100,000 entries). Entries were in base forms, with separate tables of endings. After Dictionary lookup, syntactic analysis working, first from the end of the sentence (right-to-left) and then from the beginning (left-to-right), resolved word-class ambiguities, recognised noun phrases, subjects and predicates, agreements, coordination, and resolved some semantic ambiguities. There was, however, no attempt to produce complete phrase structure analyses (the grouping of constituents evidently went no further than re cognition of noun phrase boundaries). Analysis went as far as was needed to obtain enough information to transform English structures into acceptable Vietnamese syntactic forms. A particular point was made of the mixture of syntactic and semantic information in the analysis routines. In essence the procedures were much like those of Systran at this date.

In the transfer phase, English structures were transformed into forms found in Vietnamese; thus adjective-noun groups were inverted to give Vietnamese noun-adjective groups, complex verbs were simplified ("Vietnamese tense indicators are very simple"), genitive nouns became *of*-forms (*pilot's compartment: compartment of the pilot*), and passive constructions were changed into active ones (since Vietnamese has no passive voice), e.g. *Wires can be disconnected upon removal of clip* became: *(You) can disconnect wires when you have removed clip*. The final stage replaced English lexical forms by Vietnamese.

The system produced output sentence by sentence (examples in Bruderer 1978: 308-313). The limitations of the syntax-oriented system were freely acknowledged (Byrne et al. 1970): post-editing was essential; because of the "inherent ambiguity of English" it would be possible to remove the many mistranslations only by semantic analysis or by reference to pragmatic information.

LOGOS I was the subject of a thorough evaluation by Sinaiko & Klare (1972) on the basis of an Air Force manual translated by the system in the autumn of 1970, only a few months after the public demonstration. 172 Vietnamese student pilots with knowledge of English were tested on their comprehension of the manual in the original English, a Vietnamese manual translation, the LOGOS translation unedited and the same LOGOS translation after revision. Scores for comprehension were best for the human translations, next for revised MT and worst for unedited MT; but for clarity, revised MT scored better than human translation, with unedited MT worst. The most surprising result, however, was that the Vietnamese students' comprehension scores for the English original texts were slightly higher than for the best human translations; and that, furthermore, Vietnamese who had been in the United States for five or six months did almost as well in English comprehension tests as an American control group. Similar tests with a US Navy manual confirmed the results (Sinaiko & Klare 1973). The investigators concluded that "perhaps the best way to help Vietnamese use US manuals is to improve the readability of the English text itself", which "could provide the considerable bonus of helping American users as well." The results were naturally rather unencouraging, but in any case the need for aircraft manuals in Vietnamese was shortly to come to an end.

In subsequent years the Logos Corporation continued developments of its system. By 1973 an English-Russian version of LOGOS III had been developed, and later experimental work was being pursued on systems for translating from English into French, Spanish and German (Locke 1975, Bruderer 1978). In the mid-1970s the Logos Corporation received a contract from the Iranian government to develop a multilingual system for translation into Farsi (Persian). Initially just English-Farsi, it was later to be expanded to embrace Russian, German, French, Spanish and Arabic as source languages; but, as before, Logos' plans were to be overtaken by history. No more was heard of Logos until the launch of the 'Logos Intelligent Translation System' in 1983 (Ch.13.5)

12.3: Xonics Corporation (1970-76)

A demonstration of a small-scale Russian-English MT system was given at the FBIS seminar in 1976 (Chaloupka 1976). It had been developed "in the last 6 years" by Bedrich Chaloupka (who had worked in the Georgetown project since 1956), Giuliano Grugnoli and Allen Tucker. The program was written in PL/I for an IBM 370 computer. The system could operate either in batch mode (for large volumes of text), sentence by sentence (for abstracts and titles) or interactively (for dictionary updating). There were separate SL and TL dictionaries, each containing "rudimentary" grammatical information and amounting to about 25,000 items of physics and chemistry vocabulary in either stems or full forms. No special skill or linguistic training was required to work with the dictionaries. The system was said to be "not styled on any specific linguistic theories"; it would appear to be essentially a 'direct' system of the Georgetown type, capable of "properly translating prepositions and semantic units and rearranging participle and nested structures". It is possible that experience with the Xonics system contributed to the development of the PAHO system.

12. 4: Pan American Health Organization (1976-)

The Pan American Health Organization (PAHO) is an intergovernmental agency dealing with health matters in the Americas and also serving as the World Health Organization's regional office for the western hemisphere. Its translation services deal primarily with Spanish and English, with some demand for translations into Portuguese and to a much lesser extent into French. The feasibility of developing a MT system was considered in the mid-1970s, and in 1976 PAHO contracted consultants to build an in-house system, initially for Spanish-English translation (Vasconcellos 1984, 1985). The consultants were Bedrich Chaloupka, Giuliano Gnugnoli and Allen Tucker of the TABOR company (who had worked on the 'Georgetown-type' Xonics system). At PAHO the project has been the responsibility of Muriel Vasconcellos, who in the late 1950s had been an administrative assistant in the Georgetown project (*LM* 4, 1984); in 1979 she was joined by a full-time researcher, Marjorie León.

Development of the Spanish-English system (SPANAM) continued until 1979, by which time the basic software had been written (in PL/I) and dictionaries comprising some 48,000 SL entries had been compiled. The system has been designed for an IBM mainframe computer, and (since 1979) integrated with a Wang word processor, enabling direct text input, on-line revision of MT output, and dictionary updating at the terminal. SPANAM has been operational for internal users at PAHO since summer 1980. The first major task for the new system was the PAHO biennial budget document, and by September 1983 a total of over a million words had been translated for PAHO and WHO users. A cost and quality evaluation was made on the MT output in 1981 by Macdonald and Zarechnak from Georgetown, as a result of which improvements were made to SPANAM procedures and work began on the development of an English-Spanish system (ENGSPAN).

For its MT model, not surprisingly, "the approach decided on was originally quite similar to that developed at Georgetown University in the late 1950s and early 1960s" (Vasconcellos 1984), i.e. the GAT 'direct translation' system (Ch.4.3 above) There are some differences: rather than a single SL-TL dictionary, there are two separate dictionaries linked by 'lexical numbers' assigned to a pair of SL and TL forms (and the TL dictionary is in numerical order); there are also no preliminary alphabetical sorts before dictionary searches.

The SPANAM program passes through the following stages (Vasconcellos 1984, 1985, Tucker 1984, León 1984). On input, SL words are checked against a High Frequency dictionary, then the Main SL dictionaries (with entries as either stems or full forms), and finally an idiom dictionary. The analysis program begins with a routine for homograph resolution, based on the grammatical categories of adjacent words; then, alternative translations for prepositions are considered (also according to the grammatical categories of preceding and following SL words). The next routine handles rearrangements of direct and indirect objects. Then follows a routine for handling reflexive verbs and negation (e.g. Spanish *no ha* becomes English *has not*). The next pass rearranges adjectives, conjunctions and nouns within Spanish noun phrases into appropriate English patterns (i.e. a series of SL-TL structural transformation rules). Finally, SPANAM looks up the English (TL) dictionary and a morphological synthesis routine produces TL output.

SPANAM is clearly a representative of the 'direct translation' MT model, with no separate modules of SL analysis and TL synthesis, and procedures designed specifically for the pair Spanish-English (and not vice versa). It has virtually no disambiguation (beyond distinguishing syntactic homographs), primarily because 'syntactic analysis' is limited and does not identify subjects, agents, or actions (Tucker 1984). Polysemy is therefore handled by the use of 'microglossaries' for special disciplines (Vasconcellos 1985). Any semantic analysis would require additional dictionary information beyond basic syntactic data. Post-editing is essential, and was in fact assumed and planned for from the beginning. There were no perfectionist inclinations and there are still none. SPANAM is seen as a purely practical machine aid.

Development of the English-Spanish system (ENGSPAN) began in 1982, supported by a grant from the US Agency for International Development (León 1984). Its basic dictionary has been compiled by reversing the SPANAM dictionary, amending entries and adding more English terms. It was soon realised that analysis of English demands for extensive parsing than Spanish, particularly with regard to noun phrases. For this it is intended to augment dictionary entries with information about verb valency (preferred subjects, direct objects and indirect objects, etc.) and with semantic markers (i.e. ‘human’, ‘animate’, ‘mass’, etc.) – which may also be done in the SPANAM dictionary eventually. Stages of analysis and transfer are the same as in SPANAM, i.e. dictionary lookup, verb string analysis and rearrangement, homograph resolution, noun phrase rearrangement, morphological synthesis of Spanish verbs, synthesis of Spanish nouns (León 1984). For English morphological analysis a program has been written by Macdonald. There are plans to develop an ATN-type parser for English syntactic analysis, making selective use of semantic coding, and giving improved treatment of homographs, coordination, and eventually perhaps anaphora. As with SPANAM, there is no expectation of high quality output; the project concentrates on frequent problems of English syntax and ignores fine details. It is intended also that ENGSPAN parsing will be ‘fail-safe’; some kind of translation will be produced even if part of the analysis is unsuccessful.

The PAHO system is unusual in the same organisation is involved as the developer of the MT software, as the translation service using the system, and, through its members, as the end-users of the translations produced. The level of satisfaction seems to be high: there has been a steady rise in the use of SPANAM (over half a million words translated in 1983), all types of texts have been handled (with “best performance on long technical documents and reports”), unedited MT output has been acceptable sometimes, and the post-editing rate has been high (6500 words a day). Vasconcellos concludes that “it is conservative to estimate that the gain in terms of time and cost is at least three-fold”.

Both systems[2] were demonstrated at the 1985 Georgetown conference (*LM* 25, Oct 1985), and delegates were apparently impressed by their ability to handle all types of texts, even those falling outside the fields of medicine and public health. Some evidence of the quality of Spanish-English translations is to be found in the following ‘raw’ output (from Vasconcellos 1985):

> The extension of the coverage of the health services to the underserved or not served population of the countries of the Region was the central goal of the Ten-year Plan and proba bly that of greater scope and transcendence. Almost all the countries formulated the purpose of extending the coverage although could be appreciated a diversity of approaches for its attack, which is understandable in view of the different national policies that had acted in the configuration of the health systems of each one of the countries.

12. 5: University of Saskatchewan (1965-72)

One of the projects set up by the National Research Council of Canada in the late 1960s (Ch.9.1) was the English-French MT project at the University of Saskatchewan under Kathleen Booth, who had previously worked on MT at Birkbeck College with her husband, A.D. Booth. This ‘direct translation’ system consisted essentially of a bilingual dictionary compiled on the basis of statistical analyses on a 20,000-word corpus (from the *Canada Year Book* 1962). From this were determined the most probable grammatical categories for English entries and their most frequent French equivalents (K.Booth 1967; K.Booth 1970). The stages of the system were: dictionary lookup (using Booth’s binary cut method, Ch.5.1); identification of categories and assignment of French forms; and ‘translation’ (rearrangement of verb phrase sequences, inflection of nouns, adjective-noun inversion, and output). No attempt was made to select from alternatives, e.g. *from* was given as *de/depuis/d’après*; and coordinate structures of the form ‘Adj N and N’ were analysed always (on the basis of probabilities) as ‘(Adj N) and N’ and never as ‘Adj (N and N)’ (K.Booth et al. 1971). An evaluation of the system was conducted in 1970 in which ‘post-editors’ with no knowledge of English were asked to correct 39 sentences of French output. An example extract (from Booth et al. 1971):

[2] For later descriptions of the SPANAM and ENGSPAN systems see: M.Vasconcellos and M. León ‘SPANAM and ENGSPAN: machine translation at the Pan American Health Organization’, *Machine translation systems*, ed. J. Slocum (Cambridge: Cambridge University Press, 1988), 187-235.

> L'AMERIQUE DU NORD COMPREND SIX R1EGIONS
> PRINCIPALES NATURELLES/PROPRE/ QUI SONT 2A LA FOIS/
> AUSSI BIEN QUE/PHYIOGRAPHIQUES ET GEOLOGIQUES PARCE
> QUE LES 3AGES, LES SORTIES ET LES STRUCTURES DES
> ROCHES SOUS JACENTES/FONDAMENTAL/ D1ELIMITENT/
> D1ECIDER/ LES TERRAINS/ NATURE/DES TERRES DE SURFACE.

Despite the obvious shortcomings, it was felt, nevertheless, that reasonable results were feasible after post-editing by specialists with access to the original English text. After Booth moved to Lakehead University, Thunder Bay, in 1972 no further research seems to have been done.

12.6: Other direct translation systems.

Two other, mainly short-lived, projects to produce 'direct translation' type systems are known. It is very well possible that others have also been constructed during the period.

During the early 1970s there was another English-Vietnamese MT project at the Xyzyx Information Corporation at Canoga Park, Menlo, California. The system had apparently been first developed for English-French translation of aeronautics texts on an IBM 360. The MT stages were apparently much the same as in Logos and the 1970s versions of Systran. Bruderer (1978) reports that in 1975 the dictionary comprised just 12,500 English entries; considerable post-editing was necessary. It was claimed to have been in use at some time in Canada.

An experimental program for Russian-English MT was written in FORTRAN by T.D. Crawford at University College Cardiff (Crawford 1976). Analysis was evidently restricted to phrase structure only; no resolution of homography being attempted. The BABEL system apparently produced some translations for internal use, employing a small dictionary of some 17,500 words. The project ended in 1977.