10
Systran

As a system with over 20 years of operational service, a period far longer than any other MT system, Systran cannot be ignored by any book devoted to MT. It is also a good example of the 'direct' approach to MT design, even if in certain respects a number of modifications make it less typical of this approach than older projects (see Chapter 1) and than a number of current low-priced personal computer systems.

10.1 Historical background

The genesis of the Systran system can be traced back to the earliest efforts in MT during the late 1950s. Its designer, Peter Toma, was the principal programmer of the SERNA implementation of the Georgetown University GAT system for Russian–English translation, demonstrated in 1960. (The subsequent Georgetown systems installed in 1963 and 1964 at Euratom in Ispra (Italy) and at the Oak Ridge National Laboratory of the US Atomic Energy Commission were later computational implementations of the GAT programs.) Toma had in fact begun MT research in 1957 at the California Institute of Technology before joining the Georgetown University project, the largest in the United States during the first decades of MT research (section 1.3). He left Georgetown in 1962 to set up his own company and to continue work on Russian–English MT, producing two systems AUTOTRAN and TECHNOTRAN for the fields of atomic energy and medicine. In 1964 Toma moved to Germany, where with the support of the Deutsche Forschungsgemeinschaft he began development of the Systran Russian–English

system. At this stage, the prototype, showing clear signs of its GAT-SERNA ancestry, was evaluated at the University of the Saarland with a view to adaptation as a system for Russian–German translation; however, eventually the Saarbrücken group decided to develop their own system SUSY (Chapter 11).

In 1968 Toma founded Latsec Inc. in La Jolla, California, to develop the Russian–English system for the United States Air Force (USAF). Systran (an acronym for '*Sys*tem *tran*slation') was tested in early 1969 at the Wright-Patterson Air Force Base (Dayton, Ohio) and since July 1970 has continued to provide translations for the USAF's Foreign Technology Division to this day. Subsequently, Systran was used by NASA during the joint US–USSR Apollo–Soyuz space project (1974–75), and in 1976 it replaced the Georgetown system at Euratom.

The most significant development, however, was the demonstration in June 1975 of a prototype English–French version of Systran to representatives of the Commission of the European Communities (CEC), as a result of which a contract was concluded to develop versions for translation between languages of the European Communities. The contract with Toma's new company, the World Translation Center (WTC), included an agreement for substantial development of the systems by staff of the CEC's translation department. Work began in February 1976 on the English–French version, followed by the development of systems for French–English and English–Italian. By March 1981 it was considered that each system was producing reasonable enough output to set up a pilot production service in Luxembourg. Since this date, the CEC has developed Systran systems for a number of other language pairs which are in operational use.

As well as in the United States, numerous companies were set up to promote and develop Systran, e.g. the Systran Institut in Germany, the World Translation Corporation in Canada, and the Systran Corporation of Japan. The latter developed systems for Japanese–English and English–Japanese translation during the 1980s. In addition agreements were made with various other organizations, particularly with users for the joint development of lexical databases. The complex situation was simplified, after a series of negotiations spread over a number of years, when the Gachot company in France acquired all the US and European companies involved. Since 1986 the only company with Systran rights which remains outside is the IONA company which owns the Systran Corporation of Japan and the rights to the Japanese programs.

There are now many large users of Systran (section 10.5 below) and the number of language pairs grows year by year. Table 10.1 shows the language combinations available and in development at the time of writing. In addition, from its central computer at Soisy-sous-Montmorency (near Paris), the Gachot company offers a translation service for many of these pairs and some are available on the Minitel network in France: the first MT system for use by the general public.

A sign of the maturity of Systran was the World Systran Conference in February 1986 organized by the Commission of the European Communities and held in Luxembourg. This conference — the only one so far dedicated to a single MT system — brought together all the main Systran users to exchange experience and to discuss future developments.

Available	*Under development*
English ⇔ French	English ⇔ Chinese
English ⇔ German	English ⇔ Korean
English ⇔ Japanese	English → Arabic
English ⇔ Russian	English → Danish
English ⇔ Spanish	English → Dutch
English → Italian	English → Finnish
English → Portuguese	English → Norwegian
German → French	English → Swedish
German → Italian	French → Dutch
German → Spanish	French → German
	French → Italian
	Italian → English
	Portuguese → English

Table 10.1 Systran language pairs

10.2 The basic system

The ability of Systran developers to produce a wide range of language pairs points to a relatively high degree of modularity in the system design. It is reflected in the following features. There are two main types of programs: (a) system programs, written in assembler code, which are independent of particular languages; these are control and utility programs, such as those responsible for dictionary look-up routines; and (b) translation programs which are broken down into a number of stages, each with separate program modules. Translation programs for analysis and generation are claimed to be independent of particular language pairs, the analysis module for a particular language is constant whatever the target language concerned and the generation modules are likewise constant whatever the source. Furthermore, a common Romance language analysis 'trunk' has been developed which can be implemented whenever a Romance language (French, Italian, Spanish, Portuguese) is the source in a system. This modularity has been largely achieved since Gachot acquired Systran, and it also enables the relatively straightforward introduction of new techniques wherever they seem appropriate.

Nevertheless, in certain respects, the basic procedures of Systran remain much as first conceived for the USAF Russian–English system. The main component remains the large bilingual dictionaries containing not only lexical equivalences but also grammatical and semantic information used during analysis and generation. Much of this information is in the form of algorithms to be invoked during various stages of the translation processing. While programs of structural analysis and

generation are largely independent, the main translation processes are driven by bilingual dictionaries of considerable complexity. Nevertheless, the compilation of bilingual dictionaries for new versions of Systran does not always have to start from scratch, as in many cases there need be only minor differences in the coding of source lexical items when coupled with new target languages. This fact has recently encouraged the compilation of dictionaries containing equivalents for a number of different target languages ('multi-target').

10.2.1 Dictionaries

The lexical databases for Systran are divided into the Main Stem dictionary, a bilingual dictionary of single-word entries, and various multi-word 'contextual' dictionaries. In the Main Stem dictionary every source language word (in its root form, except for English where full forms are given) is given a complete morphological, syntactic and semantic description: grammatical category, government, valency, agreement, transitivity, noun type ('animate', 'countable', 'abstract', etc.), semantic markers ('physical property', 'container', 'device', 'food product', etc.); and also a translation of the base form into an equivalent target word, accompanied by the grammatical information needed for its generation. A distinction is made between homographs with different grammatical categories, which have individual entries, and homographs of the same category, which are handled like polysemes by the 'contextual dictionaries' to be described below. This distinction is a reflection of Systran's predominantly syntax-based approach to translation, i.e. first, syntactic problems are dealt with, and then semantic information is invoked to resolve residual problems. It should be noted that for each source entry only a single target equivalent is given: this is in effect the 'default' translation which remains if it has not been changed by other dictionaries; for example, the English *station* has the default French translation *poste*.

The Systran 'contextual' dictionaries are derived automatically from a single source dictionary which can be regularly updated. They provide the data to enable analyses or translations to be modified according to context, and form a battery of dictionaries which apply at various stages of analysis and translation.

(a) The 'Idiom' dictionary is designed to deal with invariant (fixed) expressions (e.g. *on the one hand, in order to*), which may in some cases correspond to a single target form.

(b) The 'Limited Semantics' dictionary defines the scope of syntactic relations within noun phrases, e.g. by identifying *hydraulic brake* as a lexical unit and thus preventing an analysis of *hydraulic brake fluid* in which *hydraulic* modifies *fluid*. Other compounds may be identified as lexical units to ensure uniform translation (e.g. *machine translation* as *traduction automatique* and not *traduction de machine*). It also includes therefore entries for noun phrases which form a single semantic unit, e.g. French *pomme de terre* (corresponding to the single English word *potato*.) In some versions of Systran these two functions are divided between different dictionaries. The 'limited semantics' dictionary is also used to ensure the interpretations of otherwise ambiguous strings as noun phrases, e.g. *equipment cooling* as 'cooling of equipment', rather than as a noun with following modifier: 'equipment which is cooling ...'.

(c) The 'Homograph' dictionary lists the syntactic contextual information required for the resolution of certain homographs. For example in French, between a transitive verb (e.g. *prendre*) and its direct noun object (e.g. *chapeau*) there is normally a determiner (e.g. article *un* or possessive pronoun *son*), but there are exceptions (e.g. *prendre note* 'make a note of') and these are indicated in the homograph dictionary.

(d) The 'Analytic' dictionaries contain the exceptions to general syntactic rules which apply to particular words. For example, English *nor* breaks the 'normal' rule for conjunctions in that it can be followed by an inverted subject noun and verb, as in ... *nor could he see the difficulties*. These dictionaries may operate at various stages of analysis.

(e) The Conditional Semantics dictionary intervenes at the stage of transfer to make the final target language lexical selection. It incorporates both syntactic and semantic information to distinguish between potential target equivalents. For example, the default translation of English *grow* in French is *grandir*, but with an 'animate' complement it is to be *élever* and with a 'plant' as object it is *cultiver*. In some cases the number of contextual specifications is large: 400 entries distinguish the translations *huile* and *pétrole* for English *oil*.

10.2.2 Computational aspects

As stated earlier, there are basically two types of programs in Systran. The 'systems programs' written in assembler code are independent of the languages involved and include the general programs for input, dictionary look-up procedures and control of the translation processes. The 'translation programs' differ according to the specific languages being treated. As we shall see, these are divided into programs for the analysis of a source language, for transfer and for generation. These programs are written in a higher-level 'macro-language'.

Systran uses a linear data structure, comprising a sequence of records ('byte areas'), one for each word in a sentence. Each byte area consists of the word itself and the grammatical information and translation equivalents associated with the word in dictionary entries. By convention each byte stores a particular type of information. For example, byte 1 indicates the primary category (verb, noun, adverb) of the word, byte 2 the person and number of verbs, byte 3 the surface case (nominative, accusative, etc.) of nouns or the tense, mood and voice of verbs, byte 4 the gender and number of nouns, and so forth. The attributes of specific bytes vary relative to the values of other bytes (e.g. byte 3 according to whether byte 1 is 'noun' or 'verb'), but the essential feature to note is that Systran routines point to information stored in fixed byte locations. A rule to deal with a 'limited semantics' expression, for example *current practice*, might be coded in the 'macro-language' as in (1).

(1) `CURRENT $C-B26 PRACTICE (PW)`

Byte 26 specifies the adjectival modifier. In this case the rule applies if byte 26 of the principal word (PW) *practice* points to *current*. Likewise, the rule in (2) succeeds if the word which is pointed at by byte 102 has the 'semantic feature' ATTACH, i.e. *remove* must have a direct object such as *clamp*, *bolt*, etc.

(2) `REMOVE $C-B102 ATTACH`

Other bytes may trigger the initiation of subroutines, e.g. a positive value in a 'homograph notification byte' would start a homograph resolution routine (see 10.2.3 below).

During syntactic analysis, pointers are set up between specific bytes of different words, e.g. the link between an adjective and its governing noun is recorded by a pointer from byte 16 of the adjective to the address of the noun and by a pointer from byte 26 of the noun to the address of the modifying adjective. Similarly, after a subject noun has been identified, pointers are set up between it and the first word of the predicate, using specific bytes in the words concerned.

As this brief outline indicates, the data structure and low-level programming of Systran continue to reflect the computational practices of the 1960s and 1970s and are typical of many 'first generation' MT systems. Nevertheless, they have not prevented modifications and developments in a number of respects, as we shall see.

10.2.3 Translation processes

The basic stages of translation are: Preprocessing, Analysis, Transfer and Synthesis (Figure 10.1). It will become evident that the uses of the terms 'analysis', 'transfer' and 'synthesis' differ in certain respects from those found in other systems and from the usages adopted in this book.

The first stages involve dictionary look-up and morphological analysis and apply to the whole text.

1. Input: a program loads the text and identifies formatting information, e.g. titles, paragraphs, indentation.

2. Idiom dictionary look-up: invariant forms (fixed expressions) are identified, and single grammatical categories assigned (e.g. *in order to* as a preposition)

3. Main dictionary look-up: the remaining words of the text are searched for in the main stem dictionary and the information is copied into the byte areas of the data structure.

4. Morphological analysis is activated during dictionary look-up when applicable. If English is the source language, there is no morphological analysis as the dictionaries contain full forms; but in the case of languages like Russian or French stems and endings are entered in the dictionary separately. Morphological analysis entails the identification of potential combinations of stems and endings. It may also be applied to any words not found in the main dictionary in order to infer grammatical and category information.

5. Compound nouns are identified, by access to information from the 'Limited Semantics' dictionary; note that items which occur in the Limited Semantics dictionary are always treated as compounds. This causes problems when two words which otherwise form a compound happen to occur in sequence, for example the compound *femme de ménage* ('charlady') might occur in (3a) where the intended meaning is (3b).

(3a) *Il parla à la femme de ménage.*

(3b) He spoke to the woman about housekeeping.

The next stages of Analysis are applied to each sentence in turn:

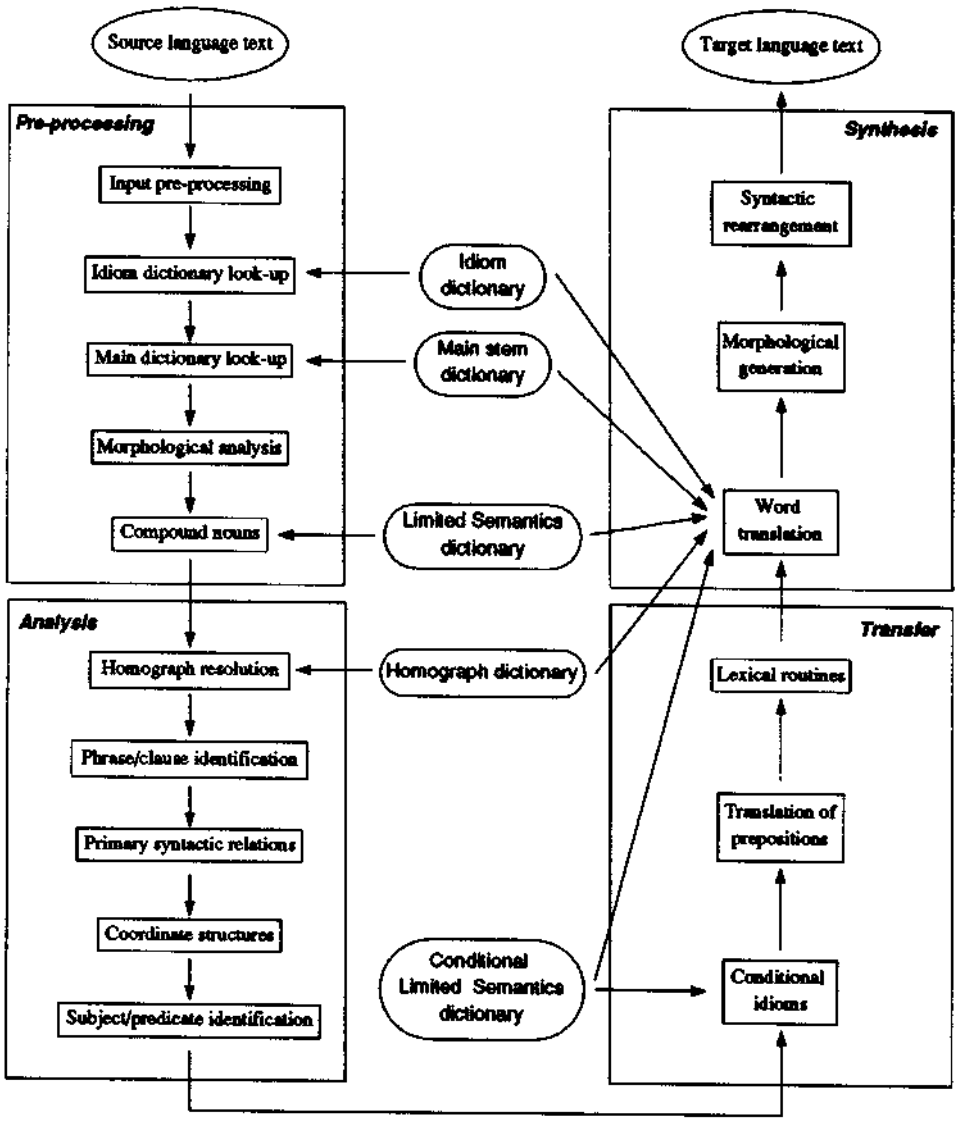

Figure 10.1 Systran translation process

6. Resolution of homographs is achieved by examining the grammatical categories of adjacent words in a single pass of each sentence. This is particularly complex in the case of English, since each word with potentially more than one grammatical category (i.e. most English words) has to be checked for its grammatical context in the string; e.g. *states* could be a verb in third person singular or a plural noun, but in the context of a preceding adjective *many* it can only be the latter. More general rules such as the following for French are also invoked: in most cases, a French verb cannot be followed immediately by a noun; there must be an intervening article or possessive pronoun. Any exceptions (e.g.

prendre note) are, as we have seen, noted in the Homograph dictionary. In the case of any unresolved homographs, the most frequent grammatical category is assumed. Note that this 'disambiguation by near context' is quite different from the parsing typically found in more recent MT and natural language processing systems.

7. Sentences are segmented into main and subordinate clauses by searching for punctuation marks, conjunctions (e.g. *because*), relative pronouns (*that*), etc. Markers for beginnings and endings of clauses are inserted. Eleven types of subordinate clause are recognised in English.

8. The 'primary syntactic relations' are determined, such as relations between nouns and modifiers (articles and adjectives, possessive nouns), between adjectives and adverbial modifiers, between verbs and objects, between prepositions and their 'object' noun (phrases), the relations of infinitives and gerunds to other words, etc. This pass also identifies the main finite verb, tense information, negation, comparatives and superlatives.

9. Words related by 'enumerations' are identified, such as coordinate structures within phrases. For example, in (4a), the coordination of *smog* and the phrase *pollution control* must be identified; in (4b) the coordination to be identified is that of *smog* and *pollution* as modifiers of *control*.

(4a) Smog and pollution control are important factors.

(4b) Smog and pollution control is under consideration.

In this case, syntactic information is available (*are* and *is*), but in other cases semantic information is necessary: for example, in (5a) the coordination of *zinc* and *aluminium* (as types of *components*) is licensed by the sharing of the semantic marker 'chemical element'.

(5) zinc and aluminium components

A supplementary routine finalises syntactic relations, e.g. in (6) the syntactic status as 'object' of *demand* already established for *speed* in the previous stage is assigned also to *accuracy*.

(6) The task demands speed and accuracy.

10. Subject and predicate are identified, going beyond the 'surface' phrase structures to 'deeper' analyses. It is a relatively simple process: already identified finite verbs are potential predicates and nouns (or pronouns) not already identified as 'objects' are potential subjects. In this way, sentences are marked as declarative, interrogative, or imperative.

11. Deep (case) relations are identified, i.e. relations between predicates and arguments, and including therefore the identification of grammatical subjects in passive sentences as the 'logical' objects of the verb. In most versions of Systran, this stage subsumes (or is preceded by) routines for establishing relations between prepositions and their governors (e.g. noun or verb).

Both stages 10 and 11 may involve consultation of the Analytic dictionaries to deal with infrequent and exceptional structures.

The Transfer program has three parts:

12. Lexical transfer of 'conditional idioms'. Standard idioms and fixed phrases have already been dealt with during analysis via the idiom dictionary

and Limited Semantics dictionary (stages 2 and 5). At this stage, other words receive 'idiomatic' translations under certain conditions which are defined in the Conditional Semantics dictionary. These conditions presuppose structural analyses of earlier stages. For example, if *agree* is in the passive, it is translated as French *convenir*, otherwise it appears as *être d'accord*; other examples involving *grow* and *oil* have already been mentioned (section 10.2.1). The homograph *lead* is to be translated as *plomb* when identified as a 'chemical element', which may have been achieved during the recognition of coordinated structures (e.g. *steel and lead*). The contexts consulted are generally the fairly simple relations of subject to verb, adjective to noun, adverb to verb, etc.

13. Translation of prepositions not already dealt with in the preceding stage. Selection is determined by dictionary information attached to verb forms or by codes attached to governing or dependent words.

14. Structural transfer using 'lexical routines', i.e. tests specified in the dictionaries for particular words or for particular syntactic or semantic categories of words. For example, the selection of appropriate translations of *as* in French (*comme, pendent que, à mesure que, puisque*) triggers alternative selections of constructions and verb tenses. As another example, routines assigned to *expect* ensure the correct constructions in French with reflexive and subjunctive, e.g. (7).

(7a) He expects to come.

(7b) *Il s'attend à ce qu'il vienne.*

The last stage, Synthesis, has also three basic parts:

15. Assignment of the default translation (in the Main Stem dictionary) for any word not already translated during transfer; for example, English *station* may have the default French translation *poste*, which would be selected if alternatives such as *gare* have not been already selected from contextual information.

16. Morphological generation, on the basis of structural information about case, gender, number, time, etc. from earlier stages and on the basis of information (in the Main Stem dictionary) about inflections and dependency restrictions, e.g. that French verbs of motion (*partir*) are conjugated with the auxiliary *être* in the past tense instead of *avoir*, and that a dependent infinitive follows directly after some verbs (*aimer aller*) rather than the normal *de* (*proposer de sortir*).

17. Generation of target word order, e.g. rearranging the word order from an English adjective–noun sequence to a French noun–adjective sequence. In French generation, this stage also deals with elision, e.g. *le homme* → *l'homme*, and with the translation of *it* by reference to the grammatical gender (male, female) of an antecedent noun in the previous sentence (usually the last) and, in this way, selecting *il* or *elle*.

10.3 Characteristics of the system

The Systran MT design is generally characterised as based on the direct approach. It is evident, however, from the description of the stages of translation that it could almost be classified as a transfer system. It would appear to have clearly separated phases of analysis, transfer and synthesis; it would seem at first sight

that the analysis stages take no specific account of target language data and that at least part of generation is independent of the source language. It is these features which lead to the claims by the developers already described above.

It must be conceded that the basic design of Systran has changed considerably over the years since the first Russian–English system was introduced in the early 1970s. At this time, its ancestry in the Georgetown systems of the 1960s was still apparent. The analysis of Russian was influenced profoundly by the needs to generate English; sequences which could be adopted unchanged in English were not analysed in depth, and structures which are peculiar to Russian were given English-like analyses. The generation of English output was likewise dependent on information about the original Russian text. Some of this mixture can be found in the present USAF version, although efforts have been made to increase its modularity and to isolate independent modules of Russian analysis and English generation.

With the English–French system for the CEC, the Systran developers moved towards much greater degrees of modularity in order that programs designed for one language as source or target could be transferred to another version with a different target or source. As a result the programs for the analysis of English are now common to all Systran systems with this language as source.

Nevertheless, Systran cannot truly be characterised as a transfer system for a number of reasons.

First, there is no separation of linguistic data into monolingual databases for analysis and synthesis, with the bilingual databases confined to straight lexical and structural transfer. Instead, Systran retains the typically 'direct' feature of large bilingual dictionaries accessed at various stages of analysis and generation. It is, however, true that different parts of the bilingual databases are accessed and applied at each stage: during analysis only the source language information is used for morphological processing, homograph resolution and the identification of structural relations. It is also true that some essentially monolingual databases are available: the Homograph dictionary appears to contain only source language information about collocations of certain lexical items, and the Analytic dictionaries seem to be restricted to structural idiosyncrasies of source languages. In these particular respects, monolingual analysis is not oriented to specific target languages.

Second, despite the labels there is no clear separation of transfer and generation. In 'Synthesis', for example, are found the main processes of lexical transfer (stage 15) and structural transfer (stage 17). 'Transfer' is confined to routines specific to particular source or target lexical items: 'transfer ambiguities' (stage 12), problems of prepositions (stage 13), and lexically conditioned structural transformations (stage 14). The only monolingual process is that of morphological generation (stage 16), which is a feature typical of earlier direct systems (section 6.3)

Third, there is no complete analysis of sentences. Various relationships are identified: nouns are linked to modifiers (adjectives and articles), adverbs to verbs, subordinate clauses are marked, coordination is specified, subjects of verbs are identified, and so forth. Structural analysis is partial and selective; no non-terminal categories are recognised (e.g. noun phrases, verb phrases) and there is

no attempt to build a full dependency or constituency representation. There is, in brief, no linguistic model or framework which is guiding analysis. The sole acknowledgement to linguistic theory is the distinction between surface relations (identified in stages 7 to 9) and deep structure relations (in stages 10 and 11).

The relative paucity of structural analysis means that much more 'grammatical' information has to be attached to individual lexical items in the dictionary than in the case of other systems, as pointed out earlier (section 6.3). In particular, as there are no representations on which general structural transfer rules can be based, the data for structural transfer have to be contained in the coding of specific lexical items.

Fourth, it appears that analysis is blocked in some cases. Once a string has been identified as an 'idiom' or 'compound' (stages 2 and 5) it is marked as a 'translated' unit. In subsequent stages there is no further analysis of the string. For example, the recognition that *hydraulic brake* is a compound means that further constituency analysis (as adjective and noun) does not take place. It should be noted, however, that this practice is by no means confined to direct systems.

Fifth, there is clear evidence of a residual preference for 'lexicographic' solutions (typical of direct systems of the first generation). It is seen not only in the precedence of lexically-driven structural transfer (stage 14) to general structure transfer rules (stage 17), but also in the prominence given to dictionary translation (idioms and compounds). One consequence is that syntactic information is not available when it could be used. Homograph resolution (stage 6) is based entirely on the examination of preceding or following words, within specified ranges. It can take no account of clause boundaries because these are not established until later (stages 7 and 8). This strategy of translating by default (which requires only shallow analysis) followed by 'repair' where possible, is typical of the first generation approach.

Finally, there is evidence of inconsistency in the application of semantic features, again, perhaps, because of the lack of a general model or framework. In origin, the semantic markers were indicators of subject domains. They now include generalised markers (PRPHY [physical property], MATER[ial], CONTNR [container], PROF[ession]), semi-specific markers (CHCOM [chemical compound], FPROD [food product], FINAN[ce]) and specific markers (MONTH, CITY, COUNTR[y]). Their primary function is to assist in the disambiguation of 'enumerations' (stage 9) and in the selection of target items (in stages 12 and 15), e.g. PROF ensures the selection of *employer* rather than *utiliser* when translating *employ*. But semantic markers are not used for semantic analysis as such. This may explain why they are not ordered in hierarchies. The marker PROF is not subordinated to HUM[an] and to AN[imate], for example, and so each would have to be assigned to the lexical items *teacher* or *lawyer*, if it is considered necessary for transfer. As the last comment implies, there appears to be no general guidance on the assignment of semantic markers; they are applied empirically to solve particular problems, with the obvious dangers of inconsistency and arbitrariness. The *ad hoc* nature of semantic markers is particularly evident in the Russian–English system. For example, the preposition *do* is translated as *up to* if the preceding verb or noun is '+increase' and as *down to* if it is '+decrease', and the preposition *po* is translated as *along* if the following

noun is '+linear', as *over* if '+nonlinear' and as *using* if '+metal tool'. Some success in regularising and simplifying the application of semantic markers was achieved with the development of the English–French system for the CEC and this has continued with later systems developed there and by the Gachot company.

From the computational point of view, Systran is also typical of first generation systems in its choice of programming environment and its general computational approach. The data structure used is primitive, while the translation process is controlled directly by a sequence of procedures written in a relatively low level programming language. Even the (relatively recent) provision of a 'rule writing macro language' does not compensate, since the formalism does not encourage the declarative style of linguistic programming seen in later systems.

The general conclusion is that although successful efforts have been made to introduce greater degrees of homogeneity and conformity (in particular, the development of a common 'trunk' for the Romance languages), Systran still reflects the eclectic character of first generation direct MT systems. It still lacks a coherent linguistic theory at its base; many routines are designed empirically for specific problematic constructions associated with particular words in particular languages; there is little generality in lexical and structural transfer; the main burden of the translation process is carried by information contained within the large bilingual dictionaries. As a consequence, methods are inconsistent, coverage and quality are uneven, and modifications of lexical information can often have unexpected consequences.

10.4 Improvability

Since much of the success of the Systran translation programs depends on the quality of their bilingual dictionaries, it has been the enhancement of dictionary entries which has received most attention by those involved with the development of systems. For example in the CEC's English–French system, to block the erroneous translation of (8a) as (8b), semantic codes were incorporated which linked the adjective *faulty* and nouns categorised as 'devices'.

(8a) The committee discussed faulty equipment and office management.

(8b) *Le comité a étudié l'équipement et l'administration de bureau défectueux.*

As another example, to ensure the correct translation of phrases including *éviter* a routine was inserted to generate a present participle, so that (9a) is translated as (9b) and not as (9c).

(9a) *éviter que l'argent soit dépensé*

(9b) prevent the money being spent

(9c) prevent that the money be spent.

However, such is the complexity of Systran's dictionaries that special care has to be taken to ensure that an 'improvement' introduced to deal with one particular problem does not degrade performance in another part of the system: this is the 'ripple effect', see section 9.5.2. (Some of these dictionaries are now very large: the Russian–English ones contain some 350,000 entries, the English–French

and German–English nearly 150,000, and the English–German nearly 100,000 lexical items.) Experience with the USAF Russian–English system has shown that improvements from the addition of dictionary entries, homograph routines, etc. are often accompanied by degradations of output quality in other parts of the process: errors were appearing where none had occurred before. On aggregate there was progress, but it was not uniform and there were substantial losses. This is the penalty for Systran's lack of an overall linguistic theory which would give clearer guidance to researchers charged with such improvements. One answer was to take greater care; proposed changes are checked against a benchmark corpus of Russian texts (ca. 50 million words), and accepted only if there is no degradation. Improvement of the system is now a matter of 'fine tuning' and not of making large-scale modifications.

Later systems, it may be hoped, may avoid (or at least delay for a longer period) the quality degradation experienced in the USAF model. There will always be this danger if changes are introduced piecemeal with no clear guiding framework, but there is little alternative with Systran: lexical data are typically irregular, and changes have to be introduced by trial and error. Amendments to the CEC's systems are consequently done on test versions and introduced into production versions only after extensive trials.

Most improvements to Systran systems, other than expansions of lexical data, have therefore been relatively peripheral, which does not mean that they are unimportant for users, particularly post-editors. In French and Italian the minutes of meetings are conventionally recorded in the present tense, but in English the custom is to use the past. At the CEC a routine was inserted to convert tenses (present to past, perfect to pluperfect, future to conditional, etc.) and to change words such as *demain* into *the day after*, rather than the normal *tomorrow*. Another improvement was the introduction of routines to deal with some types of words not found in the dictionaries. In the earlier versions, such as the USAF Russian–English system, such words were left untranslated. At the CEC, routines were written to deal at least with those which have regular endings: some routines assigned a probable semantic marker (e.g. a French word ending in *-meter* would be coded as a device, one ending in *-ologie* or *-isme* as a branch of science), others offered provisional target language endings, so that *-ogue* (as in French *radiologue*) would be rendered *-ogist* (giving *radiologist*).

10.5 Evaluations, users, and performance

The two longest-standing users of Systran systems, the USAF and the CEC, have both undertaken thorough evaluations of quality and performance on a number of occasions. The USAF Russian–English system has been in use since 1970, producing nearly 100,000 pages of texts per year mainly for information gathering purposes. It has been estimated that less than 5% of output contains errors; the system itself flags texts for not-found words, acronyms, potentially suspect adjective–noun and noun–noun compounds, uncertainty in disambiguation, known problem words, and places where some revision of word order may be necessary. In fact only 20% of texts is flagged. Such is the overall quality of raw translations

that users are now offered direct access to the system, submitting texts from personal computer terminals, usually short extracts. The USAF has recently extended its Systran services to translation from French and German and has plans for developing a Chinese–English system as well.

The English–French system at the CEC was extensively evaluated in October 1976 and in June 1978. In both evaluations comparisons were made between CEC documents in the form of human translation after revision, 'raw' unedited MT, and revised MT, each in terms of their intelligibility, their fidelity to the original, and types of errors. The raw MT output was found to have improved between the two dates in most respects, e.g. the scores of intelligibility (clarity and comprehensibility) rose from 47% to 78%, correction rates decreased from 40% to 36%. In subsequent years further expansion and improvements to dictionaries (the main source of errors) have lowered the error rate for this version to 11% in 1987. For other systems at the CEC, the rates are higher but still encouraging: French–English 14%, German–English 21%, English–Italian 29%, English–German 30%; but others are not yet satisfactory, e.g. German–French 67%.

Use of Systran at the CEC continues to grow — although more slowly than once anticipated — not only with more translators using raw MT output as 'pre-translation' drafts and with more fully post-edited translations being produced, but also with more requests for rapid minimal post-editing for texts where lower quality is acceptable, e.g. documents read for information gathering only. It is claimed that texts can be produced for this service at a rate of up to 5 pages an hour.

The use of Systran raw output for information purposes is found at many of the large companies and organisations which have installed Systran systems or which have contracts with agencies using Systran. In fact, the Systran vendors now concentrate their promotions of systems on this use rather than on the production of texts for post-editing. Major users are General Motors of Canada, where an English–French system produces literature for the Canadian market, the NATO headquarters in Brussels, the Dornier company in Germany, the German national railways (Bundesbahn), the Nuclear Research Center (Kernforschungszentrum) in Karlsruhe, the International Atomic Energy Authority, the French company Aérospatiale, and the Xerox company. The latter, in fact, is one of the biggest users, translating English technical manuals into five languages (French, German, Italian, Spanish, Portuguese) at a rate of some 60,000 pages a year. The noteworthy feature of this application is that texts are written in a controlled language, restricted in syntax and standardised in terminology and vocabulary usage, in order to reduce problems of ambiguity and homography for the MT systems and thereby to minimise and sometimes eliminate the need for post-editing. In the Xerox multilingual environment (one source and many targets) the costs of preparing texts before input are fully justified by the much lower revision costs. As a by-product, it is claimed that the use of 'Multinational Customized English' results in better-written original documents, to the benefit of English users. In coming years, Xerox intends to develop Systran systems for translating from English into Scandinavian languages (Danish, Norwegian, Swedish and Finnish).

Finally, the Gachot company has begun to make its systems available to

smaller, casual users: in France, as already mentioned, to all Minitel subscribers; and in the United States, to anyone with a terminal linked to the mainframes at the Systran organisation at La Jolla, California. Systran has come a long way from the days when all input was on punched cards and output on computer printout had to be manually revised and then retyped. Its durability and wide usage demonstrate, above all, that MT is serving practical translation needs with success.

10.6 Sources and further reading

The literature on Systran and its users is now very large. For basic initial orientation and references see van Slype and Pigott (1979), Hutchins (1986), Wheeler (1987), and Whitelock and Kilby (1983), which is particularly valuable for its detailed treatment of the linguistic and computational processes. The most accessible descriptions by the inventor himself are Toma (1976, 1977).

For more recent information see the papers given at the World Systran Conference (1986), and by Gachot (1989) and Trabulsi (1989). The Xerox implementation is described by Elliston (1979) and by Scott (1990). For recent descriptions of the USAF setup see Bostad (1988), and for the CEC see Pigott (1989). The earlier evaluations of the CEC systems are summarised by van Slype (1979).