[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1
(München: Verlag Dokumentation, 1977)]

SYSTRAN AS A MULTILINGUAL MACHINE TRANSLATION SYSTEM

Dr. P. P. Toma

President

World Translation Center, Inc.

The need for fast and accurate automatic translation continues to increase now that almost instant access to published material in all Western languages has become possible.

Twenty years of development in the field of computer translation has clearly shown that the implementation of a production system which corresponds to present user needs requires a carefully planned, universally adaptable software system which can house all the components required for the actual translation.

Many years of work and hundreds of hours of computer time were necessary to develop and debug SYSTRAN, the only existing large-scale translation system which meets these requirements.

My discussion will include the following topics:

- The details of the SYSTRAN system, its components, their structure, functions and their interrelationships.
- The SYSTRAN system's undeveloped potential applications, especially in regard to its potential use for simultaneous translation into several different languages to eventually satisfy the needs of the European Community.
- Multiple meaning resolution; procedures for adding disambiguation devices to the system's dictionaries; use of these mechanisms to satisfy individual user requirements.
- The role of semantic and pragmatic features in multilingual SYSTRAN translation.

1. INTRODUCTION

1.1 The daily increase of scientific and technical publications requires new methods to acquaint scientists with the new material in the areas of particular interest to them. Although many scientists know one or two foreign languages, the translation problem is still the greatest barrier to fast scientific and technical development.

1.2 We are facing two problems. On the one hand, human translators cannot meet the demand for rapid and accurate translations. On the other hand, specialization has created an ever-increasing need for translators to have greater degrees of technical knowledge — or at least immediate access to comprehensive technical dictionaries. Naturally, not all translators have these expensive reference works.

1 .3 It is imperative to consider economic ways to overcome the translation problem. Machine translation turns out to be economic and money saving. According to a NASA study the computer translations carried out in supporting the APPOLO/SOYUZ space mission were considerable cheaper (including post-editing) than human translation. There will be a special report during this session concerning the economy of MT on the basis of a study which has been just completed for the European Commission.

1.4 There is no time within this relatively short presentation to discuss the historic facts which led to the first and only existing and operational third generation machine translation system. Suffice it to say that this system is a third generation descendant of the SERNA System[1], which was demonstrated at the Pentagon on June 8, 1959.

[1] SERNA is derived from the Russian words S RUSKOGO NA ANGLISKI1 ("from Russian to English").

1.5 Eight years of research and the implementation of several first and second generation systems convinced me back in 1964 that the basis for a large scale universal machine translation system had to be a carefully planned, universally adaptable software system which could house all the components of a machine translation system. During the planning stage it also became obvious that the basic software should be written in the language closest to the computer's language, i.e., assembler language.

1.6 These considerations were taken into account in SYSTRAN's development. It has been proven that translations between two language can be reversed with relative ease, but what is even more important, in view of the requirements in Europe, is that, for the first time, we have a basic system which is ideally suited for simultaneous translation into different languages. The basic tasks which the SYSTRAN universal software performs are:

1) to read in the text,
2} to break it down into individual words,
3) to look these words up in the various dictionaries,
4) to attach the dictionary codes to the individual words or expressions in the text,
5) to establish an area in the computer memory where the sentence can be analyzed,
6) to control the execution of translation programs, and
7) to prepare the translation for printing, entry onto microfiche, etc.

1.7 SYSTRAN's basic universal software system is supported by various special-purpose subsystems such as systems which create and update the various dictionaries, systems which create different types of concordances according to specifications indicated on control cards, systems which extract diagnostic material, etc.

1.8 The development, programming, debugging and full implementation of the basic software system as well as its many subsystems required many years. To debug the system and especially to assure that all of its components (the individual program modules) worked together like a clock was particularly challenging. It took hundreds of nights at computer installations with top quality programmers to achieve the fully operational status we now have.

2. SYSTRAN TRANSLATION PROCESSING

2.1 I will briefly discuss how SYSTRAN works and why the system can be regarded as the first operational third generation translation system. I will particularly stress how the potentials of this system can be developed to satisfy the translation requirements of the European Community.

2.2 The first transparency is a schematic representation of the system. On the top of the picture you can see the two most frequently used input devices, i.e., keydisk and on-line. It should be noted that other input devices can be and are being used, such as keytape, keypunch and even OCR . The text passes through the computer and the translation is either routed directly to a printer or to a photo composition device for insertion of diagrams, charts, etc.

2.3 In an on-line translation mode, which you see here, the translated sentences can be routed to a terminal, where a post-editor reads the output on a screen and

[2] OCR — Optical Character Reader.

post-edits it at the terminal, and sends it on for hardcopy printing, microfiche, etc. The post-editors' corrections, by the way, could be automatically sorted according to the type of correction and used as feedback for the linguistic personnel who can then implement the needed improvements in the dictionaries or in the translation programs.

2.4 Since no human mind can anticipate all possible occurrences or uses of words or expressions, the best way to design improvements for the system is to analyze huge amounts of concordance material. SYSTRAN's ability to create concordances supplies our linguists with sufficient corpora to examine specific semantic or syntactic phenomena in large enough environments to determine firm implementable rules.

2.5 The next transparency shows SYSTRAN's universal dictionary lookup system. As the text is read in, each word is matched against a high frequency word table. If a match is found, an offset address is attached to the text word. This offset address will give an indication of how far away the dictionary information for this particular word is from the beginning of the table. The table is available during the translation in the high speed core memory and contains all the grammar and meaning information for every high frequency word in the dictionary. The program LOADTXT assigns a serial number to every word. The low frequency words, i.e., words which were not found in the high frequency table, are sorted into the same sequence as the dictionary words, and in this way the main dictionary lookup can be carried out. The sequence numbers are later used to sort the text words back to the original text sequence.

2.6 In this illustration, you see that, for the main dictionary lookup, the stem dictionary is broken up into short, medium and long stems. This is important only if the dictionary lookup handles a highly inflected language like Russian. Unfortunately, time does not allow me to describe the sophisticated mechanism of our universal dictionary lookup program; neither can I go into how the LOADTXT program handles idioms. However, I would be more than happy to answer any questions you may have during the question-answer period.

2.7 Each section (program, routine, etc.) of the translation system is called by the control program, and loaded into the high speed core (computer memory) as it is needed. The sequential control procedure is an entirely automatic process. In other words, as soon as the text is loaded in and the control program is activated, SYSTRAN independently performs each part of the translation procedure, from dictionary lookup, through source language analysis, to target language synthesis, until a complete syntactic analysis has been achieved and the total input has been translated and printed out in hardcopy form or stored on microfiche.

2.8 These programs that you see in the transparency here, i.e., HOMOR, PASS 0, PASS 1, PASS 2, PASS 3, etc., completely analyze each sentence as it comes into the Analysis Area (a specific area which the basic system reserves for this purpose). The result of the analysis is expressed in an interlingua. The creation of this interlingua makes it possible to translate into more than one language at the same time.

2.9 In the Analysis Area, 160 bytes are reserved for each word. Some of these contain information picked up during dictionary lookup. Others are used to show the results of each Pass's operation. Thus, the information in the Analysis Area includes the syntactic and semantic data retrieved from the system's dictionaries for each word as well as the results of decisions made by the analysis programs about each word's function in the sentence being translated.

2.10 The special SYSTRAN codes which are attached to individual words and to compound entries during dictionary lookup contain extensive semantic and syntactic information about each word and its interaction potential, including broad traditional concepts such as direct and indirect object requirements as well as such fine nuances as the type of subject or object preferred by a given verb. This information is examined by SYSTRAN's "parser" in determining the proper analysis of each sentence it translates.

2.11 The common syntactic features of all the languages SYSTRAN translates are expressed by the same codes, in the same position, for each language. This ensures that only minor system modifications are necessary for the translation and subsequent synthesis of any language processed.

2.12 The SYSTRAN parser includes four major "passes". The first resolves homographs. The second examines the sentence from right to left, setting switches to remember, as it progresses from word to word, exactly what types of potential syntactic relationships are possible within each clause, given the types of words already encountered. Using these switches, this pass establishes the basic structures within the sentence (verb plus object, preposition plus object, etc.) The third pass, scanning from left to right, extends these relationships, identifying enumerated objects, appositional structures, etc.

2.13 The fourth pass, using the data accumulated by the preceding passes, including information about clause boundaries, types of main and subordinate clauses, and the range or extent of embedded clauses, is then able to execute scans with the separate clauses, looking for the subject(s) and predicate(s) of each individual clause.

2.14 Each SYSTRAN system also has hundreds of lexical routines, based on source language lexical items which require special treatment. These are executed after the four major analysis passes but before synthesis and major rearrangement of sentential constituents. The synthesis programs produce the correct target language equivalent (stem plus inflections and inserted prepositions and auxiliaries, where necessary) of each source language word. The rearrangement program insures that the translated sentence will appear in the word order preferred in the target language.

2.15 It should be noted that sometimes information is passed from one sentence to another. The system reserves a special area for this in the computer memory. Such information is particularly important for the translation of pronouns whose antecedents occur in previous sentences.

2.16 During the translation process, transformations take place at several levels. A given transformation may involve a simple expression, a phrase, a clause, or, in some instances, may stretch beyond clause boundaries. The almost unlimited transformation capability incorporated in SYSTRAN is another typical feature of third generation machine translation systems.

3. SIMULTANEOUS TRANSLATION INTO SEVERAL LANGUAGES

3.1 SYSTRAN's dictionary update system compresses the source and target language information into variable length records; the size of the records is not limited. Consequently, it is possible to attach target language meanings and codes in two, three, six or even nine languages .

3.2 Once the interlingua has been produced from analysis of the source language, a basis for generating the target language has been achieved. SYSTRAN's complete open-endedness will allow it to carry out the generation of several target language equivalents simultaneously.

3.3 As we have seen on transparency #2, during the main dictionary lookup, the source language analytic and semantic codes are separated from the target language equivalents and their codes. For SYSTRAN to translate simultaneously, it will be necessary to increase the area used for target language equivalents. How much more space is required will depend on the number of languages into which the source language is to be simultaneously translated.

3.4 SYSTRAN's LS compound structure allows the attachment of several target language equivalents to each compound. Since the principal word approach allows the unlimited attachment of compound expressions to a single entry, unlimited attachment of equivalents will not create any problem.

3.5 During the actual translation phase, the target languages would be formed almost simultaneously. Let's assume we are translating from French into German, English, Italian, Danish and Dutch. After an interlingual representation of each sentence has been achieved, i.e., after the French sentences have been completely analyzed, the programs to synthesize the target languages will be executed.

3.6 The printing of translations into different languages can take place either subsequently or side-by-side.

4. MULTIPLE MEANING RESOLUTION

4.1 One of the greatest achievements in the development of SYSTRAN was the inclusion of the capability of adding multiple meaning resolutions to the system at different levels, even in the dictionary. This feature assures that the system's capacity for resolving multiple meaning problems can be increased, with virtually no limit, without increasing the size of the programs. Moreover, translators and dictionary coders can, for the first time, directly express their solutions for multiple meaning problems and incorporate them directly into SYSTRAN.

4.2 SYSTRAN has several means for solving multiple meaning problems. One of them is semantic categorization. Since semantic codes are attached to the source words, the meaning problem is really solved at the source language level. Thus, the translation into any of several languages can fall back on the choice made at the source language level. The only exception is the translation of prepositions, in which case target language prepositional requirements and other factors influencing meaning selection can be coded on the target language equivalents.

4.3 SYSTRAN has a special disambiguation feature which can be of particular use in the case of multilingual translations: conditional translation of compounds. These are called CLS (Conditional Limited Semantics) expressions. Incorporation of this feature in a third generation system allows — for the first time — a translator or a dictionary coder to add his or her knowledge to the system by using simple, well defined rules. These rules can be coded in the dictionary to go into effect at any one of several various points during translation. The rules can be used to establish meanings of any or all constituents of a given syntactic structure, or they can be used to enforce a particular parse (i.e., to establish syntactic relationships).

4.4 With SYSTRAN there is virtually no limit to the amount of information which can be added to the dictionary entries. The information is entered on specially designed universal coding sheets. After keying, the entries are subsequently organized by special update programs into variable length fields on a master tape which can then be loaded on disk. A special edit program rejects items that have been incorrectly coded and prints out diagnostic messages explaining why the individual entries were rejected. Thus, the system protects itself from human errors in coding or key-punching .

5. <u>HANDLING SEMANTIC AND PRAGMATIC INFORMATION IN MULTILINGUAL TRANSLATION</u>

5.1 As mentioned above, SYSTRAN has several different mechanisms for handling semantic information. Each of these is indefinitely expandable.

5.2 Three hundred semantic categories have been developed for SYSTRAN. These categories were developed empirically, i.e., while the system was in actual production. There is a specific advantage to developing categories empirically rather than theoretically. In this way, it is possible to ascertain their general applicability or validity in the resolution of ambiguities in live texts. The great advantage of these semantic codes is that they are used on the source language level. In other words, they are used to resolve the ambiguity within the language from which the translation takes place. This is ideal when translating simultaneously into several different languages.

5.3 The CLS approach described earlier also acts at the source language level, but has the added ability of taking target language factors into account, since, as a dictionary element, any CLS could have target language equivalents in several languages. Thus, this feature also makes SYSTRAN an ideal vehicle for multilingual translations.

5.4 The variable length feature in the SYSTRAN dictionaries also allows the attachment of a virtually unlimited number of subroutines, in compact form, to individual dictionary entries. These subroutines go into effect to resolve particular ambiguities idiosyncratic to particular lexemes or circumstances. Again, the resolution takes place in the language from which the translation takes place, but particular meanings could be entered only at certain points in these programs when the translation is carried out into different languages simultaneously.

5.5 In addition to all the types of information mentioned above, real world or pragmatic information can be included in SYSTRAN. This information can be attached either to single words or to expressions. Again, this knowledge could be used on the source language level for interrogation before establishing any particular parse or meaning. Once a decision has been made, it can be used in translation into any number of languages.

6. CONCLUSION

6.1 For the first time, there is a system which has proven its capabilities in an operational environment. It is ready to satisfy the needs of the European Commission and the European Common Market. The SYSTRAN system has the inherent capability to translate from one language into any number of languages.

6.2 SYSTRAN needs support to extend its analysis and synthesis capabilities to handle all the languages used within the European Community. Its dictionaries are open-ended; the existing terminology in these different languages can readily be incorporated. The system is ready to gradually absorb the knowledge of the best translators and to render accurate, objective and consistent multilingual translations at a speed of 200,000 to 300,000 words per hour.