[*International Conference on the State of Machine Translation in America, Asia and Europe*. Proceedings of IAI-MT86, 20-22 August 1986, Bürgerhaus, Dudweiler]

Cay-Holger Stoll

The SYSTRAN System

SYSTRAN Institut GmbH

Entwicklungszentrum
9-11, rue J.-P. Sauvage
L-2514 Luxembourg
Tel.: (0 03 52) 43 30 65

TECHNICAL DATA FOR THE IBM VERSION

The following data gives a rough survey on space and other requirements for an installation on IBM under OS/MVS.

SYSTEM CONFIGURATION

CPU	IBM/360, /730, /4341, 30xx or compatible AMDAHL, ITEL, SIEMENS 7.XXX
General Storage	Minimum requirement: 768 KB for a SYSTRAN translation run.
Tape Drives	2 drives (1600/6250 bpi) are needed during installation, dictionary updating and data protection.
Disk Storage	All compatible disk types. One 3350 pack is sufficient.
Printer	IBM TN printer (for French and Spanish TN5 or TN5S) or IBM 3800 or the corresponding laser printer or RJE connection to a text processing system.

BASIC SOFTWARE

Operating System	IBM/MFT, MVT, MVS, VS1, BS2000, BS3000 or compatible systems.
Job Control Procedures	Integration of approximately 20 procedures into the system library or into a private procedure library.
Procedural Languages	SYSTRAN programs will be delivered as load modules. Part of the service programs need the PL/1 Optimizer Transient Routine Library. Furthermore an IBM compatible SORT is required.
Editing of Text	Free choice (e.g. ATMS, TSO EDIT, SPF) or any compatible microprocessor equipment; interfaces to special text editors can be provided for.

CPU-TIME REQUIREMENTS

As the text show a wide range of variety ad possible levels of difficulty, an exact estimate concerning CPU-time requirements is almost impossible.

Translation processing itself, excluding pre- and post-editing of the text files, takes approximately 1-2 hours of CPU-time per 300.000 words on an IBM 370/168.

Costs: ask for information at the above address.

Language Pairs:

1. Operational Language Pairs:

Source Language	Target Language
German	English
English	Arabic
English	Italian
English	Russian
English	Spanish
French	English
Russian	English

2. Developed Language Pairs:

German	French
German	Spanish
English	German
French	German

3. Language Pairs under Development:

English	Japanese
Japanese	English

L I N G U I S T I C D E S C R I P T I O N O F S Y S T R A N

1. GENERAL REMARKS

1.1 The following description is based on the current French-English system. Although some features are specific to this system, as a whole, all SYSTRAN systems are built on the same pattern.

1.2 In machine translation, one refers to source language (SL) as the language a text is translated from
and to target language (TL) as the language it is translated into.

1.3 The SYSTRAN system consists of dictionaries and algorithms.

1.3.1 Dictionaries: the dictionaries are mainly of three types.

1.3.1.1 Dictionary of single words (STEM Dictionary)

Each word is given syntactic and semantic information at the SL level and one or more meanings in the TL with syntactic information attached.

The dictionaries are by definition bilingual. However, a given word's syntactic information will always remain the same in all SYSTRAN dictionaries for a given SL regardless of the TL and vice versa.

E.g.: The French verb 'venir' will be given codes to indicate its intransitivity, inflection, type of auxiliary, possible infinitive complement, concept of motion etc. in all dictionaries with French as the SL.

The English verb 'keep' will have the same inflection code (kept, kept), for gerund (keep smiling), etc. ... whether it is the meaning of the French 'continuer' or of the German 'behalten'.

1.3.1.2 Dictionary of simple expressions

When a series of words cannot be translated as a sum of the various words, an expression is coded.

E.G.: chemin de fer = railroad
not way of iron

on the other hand = d'autre part
not sur 1'autre main

1.3.1.3 Dictionary of conditional expressions

When certain words fulfill syntactic conditions, they can receive a different meaning:

E.g.: if to pull has leg as its direct object, the expression no longer means tirer la jambe but se moquer.

1.3.2 Programs

SYSTRAN is roughly sub-divided into three main linguistic parts:

1.3.2.1 Analysis of the Source Language

This set of programs organizes the parsing of the SL text, independent from the TL. This means that the analysis of English will be the same for all systems translating from English, as the analysis of French will be common for all systems with French as the SL.

1.3.2.2 Transfer modules

These programs are specific to a given language pair and solve problems of structure and/or meaning occurring between two languages. The closer the languages, the smaller the transfer.

This part is the pivot between analysis programs and synthesis programs.

1.3.2.2 Synthesis of the Target Language

This part allows for the building-up of a correct sentence in the TL, with proper endings, adequate verbal form and specific word order.

All those programs take decisions on the basis of grammatical, semantic and syntactic information provided by the various SYSTRAN dictionaries.

2. PROGRAM FLOW

2.1 Analysis

2.1.1 Preparation of Text for Dictionary Lookup (LOADTDCS)

2.1.1.1 SORT

The lookup of the dictionary is sequential and the average size of a SYSTRAN STEM dictionary is about 60.000 entries. Therefore, the words of a text are not looked up in the order they appear in the text, for it would be inefficient, redundant and CPU-time consuming.

The first step is thus the alphabetical sorting of all words of the text to be translated.

We avoid that way time wasted by repeatedly checking earlier parts of the dictionary for a match.

2.1.1.2 Separation of interrogative verbal forms specific to French as SL

In French, verbs in the interrogative are attached by a hyphen to their pronoun: Venez-vous? Etait-il? Marche-t-elle? They would not be recognized in such a form without LOADTDCS inserting a blank between the verbal form itself and the pronoun, so marking two independent words.

2.1.2 Main Dictionary Lookup (MDL)

The Main Dictionary Lookup scans the STEM dictionary of each word of the text, punctuation marks included, in the order determined by LOADTDCS, and writes the unmatched words in the Not-Found-Word list file.

MDL calls the following programs:

2.1.2.1 Morphology programs for verbs, nouns and adjectives

- o They consist of tables of legal endings for the various stems coded in the dictionary.

- o Each word is coded with an inflection code giving access to the tables.

- o When a character string is found to be an acceptable combination of an existing stem and a possible ending it gets an offset address of all the information concerning number, gender, person, tense, etc. ... from the tables and syntactic information from the stem codes.

  If no match is found, other programs are called to help for the resolution of not-coded words.

2.1.2.2 Hyphenated Word Routine

- o A flexibility in spelling hyphenated words exists, especially in English.

  Book-keeping can also be written bookkeeping and book keeping.

- o The commonest form will be coded in the dictionary leaving the task of recognizing other forms to an algorithm.

- o For each hyphenated word encountered, the routine will

  - suppress the hyphen

  - make one word out of the two

  - put those three new stems into the list of words to be looked up.

E.g.: air-hostess will produce airhostess
air
hostess

If air-hostess or airhostess is coded, its meaning will be taken; if not, the separate meaning of each word will be given.

2.1.2.3 Not-Found-Word Routine (NFWRTN)

- o It recognizes digits (literal numbers do not need be coded in the dictionary) and gives them the appropriate part-of-speech.

- o It compares other Not-Found-Words against a table of endings.

  When a match is found, a part-of-speech is attributed, with number and gender given to presumed nouns and adjectives, persons and tense to presumed verbs. This basic information will help parsing although these words will have no meaning attached to them.

- o All Not-Found-Words are put in a special file, the 'Not-Found-Word List' which will be printed with the translation output, for coding purposes.

2.1.2.4 SORT BACK

After lookup is completed, MDL puts all the words back into the text, in the original sequence order.

2.1.3 Sentence Display (GETSENTN)

This program takes each translation unit and puts it in turn in the Analysis Area (AA); each word will be allocated 160 bytes where information from the dictionaries and from the translation algorithms themselves will be stored.

GETSENTN also gives a sequence number to each word of the translation unit, in its byte 0, as a reference name for the programs.

2.1.4 Homograph Resolution (HOMOR)

2.1.4.1 For SYSTRAN, a homograph is a string of characters which can play different roles in a clause, ie. which can have different parts-of-speech.

E.g.:

L.I.K.E. in English can be, among others,

- a preposition: he is LIKE me
- a verb: they LIKE him
- a infinitive: she might LIKE it

1) Time FLIES LIKE an arrow
2) Fruit FLIES LIKE bananas

Before translating, and even parsing, these examples, it is indispensable that F.L.I.E.S. and L.I.K.E be identified:

as verb and preposition in the first example
as noun and verb in the second one.

2.1.4.2 The homograph resolution consists of one monitor (HOMOR) calling up various programs and subroutines.

2.1.4.2.1 It finds all the homographs of the sentence and calls the appropriate routine.

French as a set of 66 homographs categories, English 83.

2.1.4.2.2 Also, the monitor calls other routines which must be run before the actual parsing takes place.

- Semantic treatment of Not-Found-Words.

  While the NFWRTN, in MDL, attributes syntactic values to NFWords, the present routine gives semantic information according to the morphology of presumed nouns.

  E.g.:

  PSEUDOPHYSICIST will receive the information: concrete, countable, human, profession.

  MICRONEUROENTEROLOGY will be supposed to be abstract, mass, always singular, scientific discipline.

- Treatment of interrogatives sentences.

  It resolves the ambiguities due to the special structures of interrogative sentences and provides specific information for later parsing programs.

2.1.5 Lookup of Dictionary of expressions

*** can also be done before Homograph resolution.

The basic meaning will be replaced by a new meaning for each expression matched.

The fact that words belong to the same expression will often give clue to the later parsing programs.

2.1.6 Actual Parsing: the Structural passes.

Now, the actual parsing of the sentence can start. It will be executed in five major steps:

2.1.6.1 Definition of the clause boundaries.

This must be done at an early stage in order to establish syntactic relationships within the clause as well as to determine the limits of the search for subject/predicate.

o Each clause type is indicated in each word which belongs to it.

o Embedded clauses are recognized and indicated **so** that they can be skipped, if necessary.

2.1.6.1 Establishment of basic syntactic relationships between words.

o It will set reciprocal relationships between preposition and object.

verb and object/predicate complement

verb and adverb

noun and modifiers (adjective, article, pronoun adjective, past participle ...)

infinitive and its government.

o It also puts information on negation, auxiliary, article, comparison, etc. ... into the word modified rather than translating these indicators themselves. The TL may well express that information in a completely different way.

E.g.:

IL A MANGE	is not	HE HAS EATEN
IL NE VIENT PAS	is not	HE COMES NOT
LA PHYSIQUE EST PLUS FACILE	is not	THE PHYSICS IS MORE EASY

2.1.6.3 Establishment of enumerations

o This program finds words in enumeration on the basis of coordinate conjunctions and commas.

o It gives the same role to each number of an enumeration in the clause.

E.g.:

2 adverbs will modify the same verb

2 adjectives will modify the same noun

2 nouns will be the object of the same verb.

2.1.6.4 Search for the main subject a predicate

o It first searches for the first predicate of a clause.

o Thus, it searches for its first subject. It must be a noun or a pronoun which does not function as a complement of any sort.

o Main predicate and main subject are indicated in each word of the clause.

2.1.6.5 Establishment of deep relationships and preposition government

o Syntactic relationships (direct object, subject, predicate adjective, et. ...) are generalized into logical relationships:

action
agent of an action
object of an action.

E.g.:

THE CAT EATS THE MOUSE
THE CAT EATING THE MOUSE
THE MOUSE WAS IMMEDIATELY EATEN BY THE CAT
THE LITTLE MOUSE EATEN BY THE CAT

have all an identical deep representation:

action = to eat
agent = the cat
object = the mouse.

In this way the same deep configuration can be expressed in the structure stylistically most appropriate for the TARGET LANGUAGE.

o It decides in a subroutine on the number and/or gender of ambiguous words getting clues from the context (non-ambiguous modifiers ...).

E.g.:

GAZ is either singular or plural
LIVRE is either feminine or masculine.

but

LE GAZ LES GAZ LA LIVRE ANGLAISE LE BEAU LIVRE
are no longer ambiguous.

0 Finally, in a special routine, it finds the governing word of prepositions, basing itself on words coded in the STEM dictionary as having a strong affinity for a given preposition.

Now, the analysis of the SL is achieved.

2.2 Transfer

2.2.1 Lookup of the dictionary of conditional expressions

New meanings replace the basic meaning of certain words.

2.2.2 Preposition Translation program

- Prepositions, like all other words, have a basic meaning in STEM dictionary.

  E.g.:

  OF = DE FOR = POUR WITH = AVEC

  However, the basic meaning is often inappropriate when in specific cases of government or complement:

  IN = DANS ON = SUR

  but

  IN + name of language = EN not DANS
  DEPEND + ON = DEPENDRE DE not SUR

  These nuances are language pair dependent and are coded on the STEM entry of the governing word (or the object).

2.2.3 Lexical routines

- A monitor (LEXICAL) calls various lexical subroutines according to tables of words or categories of words requiring special treatment in a given language pair.

- The treatment might concern only a meaning (the various ways of translating the French EN into English, for instance) or a whole structure.

  E.g.:

  the French structure

  IL Y A is never correct in English as HE HAS THERE.

Sometimes, it must be translated as THERE IS... THERE ARE...

Sometimes, it will follow a time period, like 3 YEARS AGO.

More subroutines can be added when fresh problems occur between two given languages.

2.3 Synthesis programs of the target language

2.3.1 Synthesis program

The synthesis program restructures, builds-up and synthesizes the sentence according to TL rules.

- It decides on tense, person, voice of the verbal form, which may well be completely different from the SL.
- It decides on the article of a noun in the TL.
- It adapts infinitive complements to the TL: some languages might require gerunds or noun forms.

  E.g.:

  IL CONTINUE A MANGER = HE GOES ON EATING

- Inflected forms (verbs, adjectives, nouns) are constructed by attaching appropriate endings, (according to tables of endings) to the meaning stem.

2.3.2 Rearrangement of words in the TL order.

The word order is dependant on strict rules.

The rearrangement program can be highly sophisticated or practically non-existant according to whether two languages are close sisters or from different linguistic families.

French into Italian would require no special rearrangement.

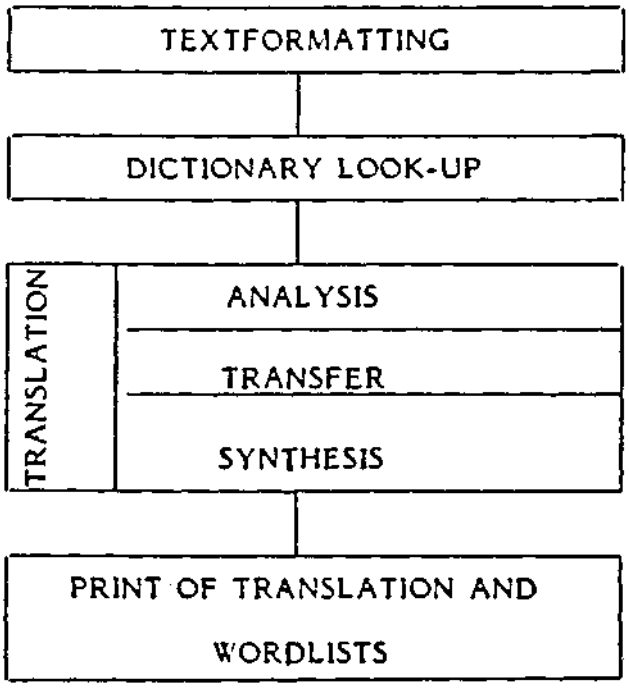
French into German or into English, on the contrary, requires numerous rearrangement routines.

E.g.:

LA VOITURE VERT	=	THE GREEN CAR
LA RESERVOIR D'EAU	=	THE WATER TANK
ELLE VIENT SOUVENT	=	SHE OFTEN COMES

etc. ...

TRANSLATION PROCESS



THE SYSTRAN TRANSLATION SYSTEM CONTAINS
THE FOLLOWING COMPONENTS