[*The Incorporated Linguist*, vol. 18, no.3, Summer 1979]

SYSTRAN

EVALUATION OF THE 1978 VERSION OF THE SYSTRAN ENGLISH-FRENCH AUTOMATIC SYSTEM OF THE COMMISSION OF THE EUROPEAN COMMUNITIES

Georges Van Slype

Director, Bureau Marcel van Dijk
Lecturer at the University of Brussels
Consultant to the Commission of the European Communities

Author's summary
Having acquired the SYSTRAN Automatic Translation System, the Commission of the European Communities had an evaluation made of the quality, cost and development potential of the system on the basis of the English-French version.

This report contains a short description of the translation system, followed by the list of evaluation criteria. The results of the evaluation are presented (key figures: intelligibility = 78%; post-editing rate = 31 to 36%; direct cost = Bfrs 35/100 words,* and conclusions drawn as to the areas of the translation market which the system is likely to penetrate and as to the advisability of continuing improvement and development work.

Belgian franc exchange rate approx. 63/£.

1. Introduction

The SYSTRAN automatic translation system was developed by Dr Toma in the United States. It has been used by the US Air Force since 1969 for the translation of scientific and technical articles from Russian into English and was also used by NASA for Russian-English and English-Russian translations in connection with the US-USSR Apollo-Soyuz space project.

This article, which is published by kind permission of the Commission of the European Communities, is a summary of a report by the same author entitled "Second Evaluation of the SYSTRAN Automatic Translation System of the Commission of the European Communities", which is available in French or English from Mr L. Rolling, DG XIII, GEC, Jean Monnet Building, Luxembourg.

The Commission of the European Communities acquired an English-French version, delivered in 1976 (an improved version was supplied in 1978), a French-English version in 1978 and a English-Italian version in 1979.

Associated bilingual dictionaries have been compiled by the Commission in the field of agriculture and food (technical, economic and administrative aspects) for the language pairs English-French and English-Italian (45000 words or stems and 11 000 expressions at the end of 1978) and in the field of metal fabrication for the French-English pair (15 000 words or stems and expressions at the end of 1978).

The two successive versions of the English-French system have been evaluated as described in this report.

2. Brief description of the system

The basic components of the system are:

a software package

a number of bilingual dictionaries.

The dictionaries are compiled by lexicographers or coders, who are either qualified translators or have a master's degree in languages, at the rate of some 50 to 100 entries per day.

Each entry comprises:

— The stem of a word (the word without its grammatical flexional endings) in the source and target languages
— grammatical data: part of speech, gender, number, . . .
syntactical data: prepositions, parts of speech and complements governed by the word, . . .
— semantic data aimed at eliminating ambiguities due to homographs (words that can be more than one part of speech, eg "kind": noun = *sorte* or adjective = *gentil),* polysemy (words which may have a multiplicity of meanings within a given part of speech, eg "plant": *usine* or *plante)* or due to expressions which cannot be translated literally (eg "information retrieval": *recherche documentaire* and not *recouvrement d'information).*

The software, which contains some 100000 instructions, falls into two parts:

— the first, which is independent of the language pair, written in Assembler language by the system's designers
— the second, which is specific to each language pair, written in "macro-language" by linguists who are members of the system designer's or the Commission's team.

The translation procedure is as follows:

— recording of the texts in the source language on punched cards or magnetic tape
— input of the text to be translated into the computer
— automatic lookup of the dictionaries to retrieve the data relating to each word or expression in the source text
— analysis of source text, in the following sequence : recognition of expressions, solution of homograph problems, clause separation, identification of noun and verb groups, recognition of enumerations (groups separated by "and" or "or"), search for subject and predicate
— synthesis of target text, in the following sequence: translation of prepositions between governing and governed words, solution of problems of polysemy, addition of the correct endings, adapting of the word order to the syntax of the target language
— printout of target language text only or synoptic presentation of source and target language texts.

3. Evaluation of automatic translation

3.1 *Scope of evaluation*

Automatic translation has not yet achieved a breakthrough, and apart from a very few users it is still at the research and development stage.

Hence the importance of making an evaluation of its performance and potential which is as accurate as possible, since such an evaluation will in fact determine whether development work on these systems will be continued or abandoned. The ALPAC (Automatic Language Processing Advisory Committee) report provides a good illustration of the responsibility falling on the evaluator: the study was carried out between 1964 and 1966 by a group of linguists, mathematicians and computer experts who set unrealistic standards of perfection and aimed at a rigorous equivalence between the products of human and machine translation, and was based on an evaluation of first-generation systems. The ALPAC report resulted in the cutting off of finance to some twenty to thirty teams carrying out research in this field in the United States.

When, in 1975, the Commission of the European Communities became interested in automatic translation, the initiative originated in its Directorate-General for Scientific and Technical Information and Information Management, and the objectives were quite different: the motivation to see machine translation develop is obviously much greater in Europe (6 languages with equal legal status, 4 of which giving rise to a substantial volume of translation) than in the United States (a single language with a dominant position in scientific and technical publications and resulting in an extremely limited amount of translation). In addition, the potential of automatic translation is seen in Europe at the political level (contribution to breaking down linguistic barriers) and not primarily as a technical problem (achievement of performances of the same level as human translation).

3.2 *Evaluation criteria*

The setting of objective evaluation criteria is an extremely complex task: each of the parties involved in the translation process has different requirements, which themselves vary according to circumstances:

— the end user of a translation may require:
  • a meticulous piece of work reproducing all the shades of meaning in the original (examples: legal texts, literary texts)
  • a good quality translation with a high level of intelligibility and fidelity (examples: scientific, technical and administrative texts, manuals, etc)
  • a translation of good enough quality for him to understand the contents of the document without being greatly concerned by the accuracy of detail (examples: working documents, scientific, technical and administrative texts when the reader wishes to assess their relevance before possibly requesting a quality translation)
— the person correcting the translation, a highly-qualified specialist, who revises the human translation or "post-edits" the machine translation will demand a job which is commensurate with his qualifications: he will not appreciate having to correct numerous errors, many of which are of a mechanical and repetitive nature
- the head of a translation office will be concerned
  • if he runs an official department, with the interests of his staff, and he will not therefore wish to impose on them a task which is not only uninteresting (post-editing) but also unusual in the translation profession, whereas
  • if he manages a private company,

he will be more interested in the economic aspects of machine translation: more rapid service, cutting of costs, development of the market, etc.

-- the person responsible for acquiring a translation system will be concerned with its acceptability to the end user and/or the reviser, the saving in costs and the development potential of the system.

An array of criteria is required to take these different concerns into account, yet at the same time the evaluation machinery must not become cumbersome.

In February 1978, the Commission decided to organize a workshop on the methodology of evaluating machine translation, bringing together the main European and American teams engaged in research and evaluation of machine translation (*).

On the basis of the conclusions reached by this workshop and the experience gained from the first SYSTRAN evaluation, the following criteria were selected for the assessment of the 2nd version of the English-French SYSTRAN:

Macroevaluation (assessment of the system's overall performance):

- *intelligibility,* that is to say a subjective evaluation (but carried out by several evaluators) of the degree of clarity and comprehensibility of each sentence. The Carroll 8-point scale used by the ALPAC Group was not adopted: although based on valid psychometric principles (equal-appearing intervals) this scale in fact measures both intelligibility and style (example: rating of 8 = "perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or mildly unusual word usage that could, nevertheless, be easily corrected"), and therefore tends to be systematically to the disadvantage of the machine translation: many sentences translated by the computer are in fact perfectly intelligible although stylistically incorrect. We have therefore used a much more rudimentary scale, which nevertheless has the advantage of being easy to interpret and unambiguous (measures intelligibility only):
  3 very intelligible
  2 fairly intelligible
  1 barely intelligible
  0 unintelligible
- *fidelity:* subjective evaluation of the degree to which the information contained in a sentence in the original text has been reproduced without distortion in the translation (also on a 4-point scale)
- *style:* subjective evaluation of the correctness of the style of each sentence (4-point scale)
- *reading time:* time required to read and understand a text, or to realize its unintelligibility, but not to memorize it
- *revision and post-editing time:* the time taken to go through a translation, with reference to all or part of the original text, carry out terminological research and correct the translation
- *correction rate:* ratio of the number of words corrected to the number of words in the translation
- *acceptability to the end user:* subjective evaluation, by a sample of end users, of the degree to which the translation is acceptable.

Microevaluation (analysis of the errors attributable to the system and the means of avoiding them, with a view to assessing the extent to which the system can be improved):

- *source of translation errors:* detecting the probable source of each of the errors corrected by the post-editor of the machine translation
- *analysis of the main causes of errors, the appropriate remedies and their side-effects:* determination not only of the causes, but also of the remedies, the amount of work required to implement such remedies and their possible side-effects.

4. Results of the macroevaluation of quality (sample of 650 sentences containing 12 300 words)

- *Intelligibility:*
  - original text 99%
  - machine translation, without post-editing 78%
  - revised human translation 98%
  - post-edited machine translation 98%

— *Fidelity of the machine translation:* without post-editing 73%

— *Style of the machine translation:* without post-editing 76%

— *Reading speed by end user:*
  - original text 3 700 words/hour
  - machine translation, without post-editing 3 200 words/hour
  - revised human translation 5 000 words/hour
  - post-edited machine translation 4 300 words/hour

— *Translation and correction time:*
  - human translation without revision 22 min/100 words
  - revision of human translation 7 min/100 words
  - machine translation (CPU time) 0.05 min/100 words
  - post-editing:
    * by translator/revisor (*) 22 min/100 words
    * by engineer/end user (*) 8 min/100 words

— *Correction rate:*
  - revision rate for human translation 12%
  - post-editing rate for machine translation
    * by translator/reviser (*) 36%
    * by engineer/end user (*) 31%
  - type of corrections:

words replaced	53%
words corrected	26%
words transposed	9%
words deleted	6%
words added	6%
Total	100%

— *Acceptability to end user (preliminary survey conducted among 17 users in 3 institutions):*
  - 88% consider that the machine translation without post-editing is acceptable under certain circumstances
  - 71% would be interested in receiving unrevised machine translations:
    * from languages they do not know
    * for their personal documentation and for working documents
    * given a rapid service of about half a day
    * given low cost.

5. Cost of the machine translation (Calculations carried out during the first evaluation of the English-French SYSTRAN, which are reproduced here to provide all the relevant data for the conclusions drawn on the macroevaluation in § 6):

— Input Bfrs 100/100 words

— Machine translation (cost of machine time; excluding depreciation of software and dictionaries) Bfrs 35/100 words

— Post-editing (by translator in a large administration) Bfrs 300/100 words

— Typing of fair copy Bfrs 40/100 words

Total: Bfrs 475/100 words

Comments:

— the cost of human translation with revision varies from Bfrs 150 to 250/100 words (free-lance translator) to Bfrs 500/100 words

(*) The proceedings of this workshop will shortly be published by the Commission.

(•) Intelligibility of the post-edited text was 98% in both cases.

(translator and reviser in a large administration)

— if the text exists in a computer-readable form, the cost of input is replaced by the cost of a simple automatic reformatting.

— if the text is post-edited in a translation department organized specifically to carry out this type of work, the cost of post-editing can be substantially reduced;

—if the unrevised machine translation is supplied direct to the end user, the cost of the process becomes extremely low: less than the cost of typing the translated text!

—the trend in the future will be towards:

- an increase in wage costs
- a fall in or stabilization of data-processing costs.

6. Conclusions on the use of SYSTRAN

An analysis of the data presented above suggests that machine translation has not yet reached a stage where it provides a substitute for human translation. This having been said, it does appear that a system of the quality of SYSTRAN can fill two important gaps in the market:

— *pretranslation:* the supply, within a very short time and at low cost (particularly if the source texts are available on a machine-readable medium), of machine translations to translation departments; the latter will then undertake the post-editing and supply of a finished product, the quality of which will be comparable in all respects to that of a purely human translation but costing less in terms of both money and time. The creation of such a service naturally presupposes:

- the existence of specialized dictionaries in the fields covered by the translation department
- that the operation of the data-processing department permits translation requests to be dealt with virtually immediately and that there are adequate telecommunications links between the translation and data-processing departments.
- that the translation department is structured to take into account the specific nature of the post-editor's job, which is fundamentally different from that of the reviser.

— *rough translation:* the supply of unrevised machine translations to the end user, whenever the latter wants to get an idea of the contents of a document in a foreign language but does not require a perfect translation.

7. Results of the microevaluation

— *Causes of the errors corrected at the post-editing stage*

• source text (ambiguities, syntactical and spelling mistakes)	5%	5%
• input	1%	
• stylistic appreciation of the post-editor	14%	
• translation system		
dictionary	50%	
analysis	16%	
synthesis	10%	
miscellaneous	4%	
	100%	

— *Causes of and remedies for the main errors:*

This part of the microevaluation resulted in the drawing-up of a highly analytical list of the types of improvement which should be made to eliminate the main errors detected (amendments to dictionaries, grammar rules, etc) specifying, in each case, the time required and possible side effects.

8. Conclusions on the microevaluation

It would appear that the English-French SYSTRAN is still capable of being improved by a vast number of detailed refinements, each of which:

- will make a very small individual contribution to the improvement of the system; and the significance of which varies according to the nature of the amendment
- will require a more or less substantial effort in man hours by coders or linguists.

In the light of this, it would appear advisable to adopt a strategy of optimum improvement by selecting, the refinements resulting in maximum improvement at minimum cost.

9. Overall conclusion

The evaluations made of the two versions of the English-French SYSTRAN lead one to recommend the continuation of work on improving the system and making it fully operational.

Improvement of the system should be based on a policy of achieving the optimum benefit from the resources employed.

If the system is to become operational, pilot schemes must be organized in conjunction with public and private translation services and individual and institutional end users, in view of breaking down the "innovation barrier".