[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

The Respective Roles of the "Industrial Evaluators" and the System Developers in the Evaluation of MT Systems

Sylvie Régnier
AEROSPATIALE

The evaluation method we developed at the Research Centre of the French company AEROSPATIALE in 1989 results from the need to find basic criteria to evaluate the performance of an MT system **in an industrial environment.** This method is based on pragmatic considerations rather than a thorough scientific approach. The aim is to develop a methodology reusable for different MT systems and available to managers as **a decision tool** for the purchase and general use of an MT system in a large company.

We found out that one of the best ways to evaluate an MT system is to examine its ability to adapt itself to the company-specific needs. The collaboration between the industrial evaluator and the system developer seems therefore to be a crucial issue to guarantee an effective and realistic evaluation.

This evaluation approach and our various trials of different MT and CAT systems, especially SYSTRAN, ARIANE and WEIDNER, highlighted a number of limits and restrictions in the use of such systems. Our general approach, based on our many year experience in that respect, is that there is no universal Machine Translation system, i.e. no MT system will ever be able to provide acceptable raw translations of any scientific document, in any technical field and for any vocational purposes. We think that a prerequisite to an effective use of such tools is to select specific technical fields with text corpora and to identify and analyse their terminological, grammatical and syntactic characteristics. The quality level of the final post-edited translations also has to be defined prior to introducing an MT tool in a translation department.

Most of the time, however, these steps only appear to be necessary in the process of experimentation. When an industrialist purchases or trials an MT system, neither grammars nor dictionaries have been explicitly modified to meet his specific needs. The role of the evaluation is therefore to judge

whether the system, over a certain period of time, proves to be adaptable to specific industrial requirements. The evaluation method used at AEROSPATIALE therefore implies an interactive process between the industrial evaluator and the system developer.

The different steps are the following: First of all, the industrial evaluator analyses and classifies the errors recorded in raw machine translations. He/She then defines, together with the system developer, the solutions to the coding of polysemic words, user-specific terminology and grammatical or syntactic requirements. Subsequently, the system developer performs the corresponding corrections of the previously highlighted terminological and linguistic deficiencies of the MT system and provides detailed information about the work carried out as well as the reasons why a number of problems could not be solved. Finally, the industrial evaluator controls the relevance and accuracy of the modifications performed, using the technique of test suites.

The linguist in charge of MT projects at AEROSPATIALE carries out a twofold evaluation: a text based evaluation in a first phase of preparation and a test suite based evaluation as a control tool.

The method used for the text based evaluation is also a very pragmatic one. The principle is to study the results of the raw machine translation and a reference translation, which might be either a fully human translation or a post-edited one, compared to the source text. This comparative analysis is initially performed on hard copies of the three texts involved using a system of different colours to show the various classes of problems recorded. Different categories of errors have thus been identified according to the type and delay of correction they involve and a colour is used for each class.

The **terminological errors** first may be of two types: either non-pertinent terminology (highlighted in yellow) or unknown/unrecognized words (in green).

The second upper category of problems concerns the **linguistic mistakes.** They are subdivided in 4 classes: *word order* (blue), *prepositions* (pink), *grammar (pink)* and *syntactic analysis* (also pink). The same colour is used for the last three types of errors because it is often difficult, for the industrial evaluator, to determine whether a badly translated noun or verbal phrase is imputable to a deficient grammar rule or syntactic analysis.

The next step of this text based evaluation is the elaboration of a coding sheet or form describing the different mistakes identified in the first phase with their semantic environment. These coding sheets also contain information on the parameters chosen during the machine translation process (language pair and topical glossaries selected) together with economically orientated data such as the time devoted to the whole process of translation and the volume involved. Examples of filled in coding sheets are given in the annex.

In that respect, it may be noteworthy that different processes are involved in the use of Machine Translation systems and are to be taken into account when considering statistics or percentages of productivity. As a matter of fact, the handling of the source text may be a problematic issue from the beginning: depending on whether it is a hard copy or a machine readable document and whether the hard copy is a top copy, a photocopy or a telecopy, the tools needed as well as the corresponding time and cost may vary considerably; it may range from the simple reading of a diskette, to the use of an optical character reading software or the retyping of the whole text. Furthermore, the post-editing process may be tricky and time consuming if, for example, non-textual information is to be reinserted in the final document. Statistical measures on the profitability of an MT software in an industrial environment should therefore be qualified by many different criteria which mostly depend on a specific context of use.

The coding forms are then handed over to the system developer together with the three highlighted texts. They also provide a first estimation of the quality of MT to the engineer or manager who had ordered a machine translation of his/her text. Our experience proved that the approach of the general public to MT is either too optimistic or too pessimistic. In our attempt to make our colleagues sensitive to MT and aware of its opportunities and limits, it appeared necessary to provide them with such detailed coding forms rather than the simple raw machine translation, which in most cases would dissuade them from going on with experimentations.

The next step in our method is to perform a control of the work carried out by the system developer, provided he/she has sent back the required coding feedback information.

This second evaluation aims at checking that the terminological and linguistic problems previously identified have been properly solved and that the modifications apply to any context of use (test suites). The first task is to examine the changes in behaviour of encoded words, noun or verb phrases or sentences in different grammatical, syntactic and terminological environments and with various parameters (especially various topical glossaries). The second aspect of this study focuses on the possible side effects of these modifications on the remaining constituents of a phrase, sentence or paragraph.

The implementation of this method has however raised a number of problems especially when working out test suites. The first difficulty is related to exhaustiveness: it is not an easy task to identify all the possible contexts of use which may occur from one corpus to another. Secondly, the methodological problem is to define common criteria and reusable parameters for the evaluation of problem classes. At the time being, the test suite sentences we use are often created on a random basis.

The result of the second evaluation is then forwarded to the system de-

veloper for examination and possible final modifications.

The whole process is time consuming and costly for the two partners but may result in a rapid improvement and adaptation of the initial MT system. Both the industrialist and the system developer may thus benefit from such a cooperation.

Generally speaking, the major industrial requirements related to the use of MT systems are the following :

- ease of updating the user-specific terminology,
- cooperation and open exchange of information between the evaluator and the system developer,
- adaptability of the MT system to specific industrial needs and constraints and
- development of user-friendly computer-aided pre- and post-editing tools.

The last but not the least criterion to be taken into account in the use of MT systems in an industrial environment is the expected quality level of the post-edited machine translations. It clearly appears that the post-editor (who is often a skilled translator) and the end-user do not have the same estimation of this level.

For the post-editor, the difficulty is to find objective criteria of (un)acceptability of raw machine translations. He/She has to find a compromise between his/her wish to provide a high quality translation comparable to a fully human one and the necessity to take the most of the raw machine translation and to meet productivity, profitability and rapidity requirements. No formal measures seem to make a clear distinction between post-editing and rewriting or retranslating an MT output ; the threshold is very difficult to determine.

On the other hand, the end-user appears to be much more tolerant as far as the quality of the post-edited translations is concerned. Depending on the language pair involved, he has different requirements. If a document is to be translated from a foreign language to his/her mother tongue, the user often wants to obtain rapidly the rough content of the text so as to check whether it is of interest for his/her field of specialization. This initial information gathering purpose may then reveal a need for a high quality translation.

Should a document be translated from his/her mother tongue into a foreign language, the quality requirements are considerably higher. Most of the time, the resulting post-edited translations are destined to be distributed abroad, for example within the framework of international cooperation projects. This case also applies to after sales documentation such as maintenance manuals which must be regularly updated and may represent

huge amounts of documents for which any translation mistake, misunderstanding or haziness is likely to have highly serious repercussions.

As a conclusion, evaluating MT systems requires taking many different aspects into account and is often a time consuming activity. The various actors involved in Machine Translation – developers, distributors and users – also have their specific needs and constraints which have a decisive influence on their final evaluation of an MT system. Closer contacts between these three types of actors may be a way to improve the quality of MT in general and to reach a consensus on what tomorrow's MT should be.

ANNEXES

CODING SHEET N° 01

FORMATION DES POLLUANTS

I/ LANGUAGE PAIR

French --> English

II/ GLOSSARIES AND DICTIONARIES SELECTED

. Tgs : - Aviation et space
- Chemistry
- Physics
. CUST = 2

III/ MACHINE TRANSLATION PARAMETERS

. Diskette		: 2219 mots
. Machine Translation	Nb of pages	: 6
	Nb of batches	: 1

IV/ ANALYSIS OF THE ROUGH MACHINE TRANSLATION

a) Terminology	: 70
b) Unrecognized words	: 4
c) Word order	: 20
d) Prepositions	: 15
e) Grammar	: 5
f) Syntactic analysis	: 10

FICHE DE CODIFICATION N° 01

FORMATION DES POLLUANTS

A) TERMINOLOGIE

N°	TEXTE SOURCE	TRADUCTION BRUTE	TRADUCTION HUMAINE	CODAGE		
				Tq	CUST	CSD
A.1	formation (de polluants)	formation (of pollutants)	(pollutant) build-up process			
A.2	richesses locales du mélange / de fonctionnement	local richnesses of the mixture / of operation	(local) fuel-air ratios			
A.3	traduire + N abstrait	to translate	to express			
A.4	concepteur	originator	designer			
A.5	prévoir + N concret	to envisage	to predict			
A.6	(chambre) de combustion	(combustion) chamber / room	combustor			
A.7	phase de dimensionnement	phase of dimensioning	design stage			
A.8	régime de fonctionnement	operating speed	rating			
A.9	dans les conditions ralenti	under the conditions slowed down	in idle conditions			
A.10	tube à flamme	tube with flame	flame tube			

FICHE DE CODIFICATION N° 01

FORMATION DES POLLUANTS

A) TERMINOLOGIE (suite et fin)

N°	TEXTE SOURCE	TRADUCTION BRUTE	TRADUCTION HUMAINE	CODAGE		
				Tg	CUST	CSD
A.68 A.69 A.70	par effet de + N procédé injecteur (chimie) compromis ralenti / plein gaz	by effect of + N burner compromise the slowed down / full throttle	by +ing injector trade-off between low and high power settings			

B) MOTS NON RECONNUS

N°	TEXTE SOURCE	TRADUCTION BRUTE	TRADUCTION HUMAINE	CODAGE		
				Tg	CUST	CSD
B.1 B.2 B.3 B.4	Hydrocarbures Gaz Alcoôl prévaporisation	Hydrocarbures Gaz Alcool %prévaporisation	hydrocarbons gas alcohol preevaporation			

CODING SHEET N° 01
FORMATION DES POLLUANTS

C) WORD ORDER

N°	SOURCE TEXT	MT OUTPUT	REFERENCE TRANSLA°
C.1	paramètre de charge aérodynamique	aerodynamic parameter load	air loading parameter
C.2	cycle atterrissage décollage	cycle landing takeoff	Landing-Take-Off cycle
C.3	on constate en général	one in general notes	it can generally be noted
C.4	X diminue en général	X decreases in general	X generally decreases
C.5	méthodes de prévision globales /semi globales	methods of forecast total / semi total	global and semi global prediction methods
C.6	mieux orienter qch	to direct sth better	to better orientate sth
C.7	X où se produit Y	X where occurs Y	X where Y occurs
C.8	calcul de champ thermique	thermal calculation of field	velocity field calculation
C.9	les phénomènes physiques mis en jeu	the brought into play physical phenomena	the physical phenomena involved
C.10	les ressources informatiques nécessaires	the data-processing resources necessary	the required computer means
C.11	normes internationales .OACI	international standards OACI	ICAO international standards
C.12	ils vont également augmenter	they also will increase	they will also increase
C.13	au niveau des futurs projets de	on the project level future of	for the future projects of
C.14	X sur lequel doit porter l'effort de	X on which must carry the effort	X on which the effort must focus
C.15	diverses techniques particulières	various techniques particular	various specific techniques
C.16	turbines industrielles .THM	industrial turbines THM	THM industrial turbines
C.17	températures de flamme plus basses	temperatures of lower flames	lower flame temperatures
C.18	système (d'alimentation en carburant) plus complexe	(feeding) system (while carburizing) more complex	more complex (fuel supply) system
C.19	cela ne donne donc pas ...	it thus does not give	therefore it does not give
C.20	permettre de limiter ainsi qch	to make it possible to thus limit	to allow thus to limit

CODING SHEET N° 01
FORMATION DES POLLUANTS

D) PREPOSITIONS

N°	SOURCE TEXT	MT OUTPUT	REFERENCE TRANSLA°
	à / au		
D.1	à un régime de fonctionnement / élevé	with an operating speed / to a high mode	at a (high) rating
D.2	tube à flamme	tube with flame	flame tube
D.3	zones à fortes températures	areas at strong temperatures	high temperature zones
D.4	à la .SNECMA (nom de société)	with / to the SNECMA	at SNECMA
D.5	(hétérogénéités) internes à qch	internal (heterogeneities) with / to sth	sth internal (heterogeneities) / specific to sth
D.6	à X donné, S + V	with X given,	for a given X,
	à même X, S + V	with / to same X,	for the same X,
D.7	consister à + inf.	to consist of +ing	to consist in +ing
D.8	conduire à X et donc à Y	to lead to X and thus with Y	to lead to X and therefore to Y
D.9	au ralenti	to the idle	at idle
	de / des		
D.10	produits intermédiaires de la combustion	intermediate products of the combustion	intermediate products (originating) from the combustion
D.11	diminution des fumées / accroissement des niveaux	reduction in the fume / increase in levels	reduction of smoke / increase of levels
D.12	De l'analyse de X, il apparaît que	Analysis of X, it appears that	Analyzing X evidences that
	dès		
D.13	dès la phase de dimensionnement	as of the phase of dimensioning	as from the design stage
	comme		
D.14	considérer qch comme + adj	to consider sth like	to consider sth as
	par		
D.15	optimiser qch par le calcul	to optimize sth by calculation	to optimize sth through calculat°

CODING SHEET N° 01
FORMATION DES POLLUANTS

E) GRAMMAR

N°	SOURCE TEXT	MT OUTPUT	REFERENCE TRANSLA°
E.1	ce sont en fait les X qui pilotent qch	it is in fact the X which controls sth	sth is actually governed by X
E.2	les zones où +S+V	the areas when	the zones where
E.3	ils ont adopté X comme réglementation	they adopted X like regulation	they have adopted X as regulations
E.4	le débit de réducteur à injecter	the flow of reducer to inject	the flow of the reducing agent to be injected
E.5	l'évolution de X depuis n décennies peut être constatée ...	the evolution X for n decades could have been noted ...	the evolution of X over n decades is illustrated...

F) SYNTACTIC ANALYSIS

N°	SOURCE TEXT	MT OUTPUT	REFERENCE TRANSLA°
F.1	c'est dans les phases X que +V+S	it in the phases X which + V + S	S + V in phases X
F.2	la phase "montée"	the phase "is installed"	the "climb" phase
F.3	des températures qui sont encore élevées	temperatures which are still raised	temperatures which are still high
F.4	X l'emporte sur Y	X carries it on Y	X prevails over Y
F.5	adv., dès + N des méthodes + V	adv., as of + N of the methods + V	adv., as from +N, methods + V

CODING SHEET N° 01
FORMATION DES POLLUANTS

F) SYNTACTIC ANALYSIS (ctd)

N°	SOURCE TEXT	MT OUTPUT	REFERENCE TRANSLA°
F.6	X donne, à même Y des niveaux de	X gives, with same Y of the levels	X generates levels for the same Y
F.7	X a entraîné par effet de Y une diminution	X drove by effect of Y a reduction	X resulted in a reduction by Ying
F.8	système d'alimentation en carburant	feeding system while carburizing	fuel supply system
F.9	l'évolution des longueurs de chambres des moteurs	the evolution lengths of rooms of the engines	the evolution of engine combustor lengths
F.10	ceci a permis d'obtenir sur des moteurs modernes des niveaux ...	it made it possible to obtain on the modern engines of the levels ...	on modern engines, it permitted to obtain levels ...