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

A procedure for the evaluation and improvement of an MT system by the end-user

Brigitte ROUDAUD
B'VITAL (groupe SITE)

1 Introduction

SITE and B'VITAL are currently working on a project which aims at commercializing the French-English CAT system, developed by B'VITAL. This system is based on ARIANE, GETA's machine translation system.

GETA is a university research laboratory in Grenoble, France, which has been involved in machine translation for 30 years. The first version of ARIANE, inaugurated in 1978, was in fact the successor to a previous system, CETA.

One of the main goals of the project is to validate the system, in terms of efficiency. It is difficult to define exactly what *efficiency* is for a CAT system, as the use of such systems is not yet completely defined. For a commercial company such as SITE which specialises in technical communication, the CAT system will be one of the tools at the translators' disposal. It should be integrated as far as possible into the whole documentation process and should provide useful help in saving time or in increasing the quality of the document.

Another important aspect of the system's evaluation is its up-grading capacity. At present no CAT system is perfect (what does *perfect* mean, anyway?), but how far they can be improved is an interesting issue for the end-user. Evaluation of this kind must be done in collaboration with the linguists, who are responsible for the grammar development, and the lexicographers, who are responsible for the dictionary build-up.

The validation procedure will consist of an *in-house ß-test.* The company's translators will use the system to translate existing documentation and will then revise the raw translation. To evaluate the up-grading capacity, the revisers will make known their diagnoses, using what we call *problem sheets.*

During the ß-test, the system will translate 5000 pages of aeronautical documentation. The documentation has been chosen for the following reasons:

- it is reasonably representative, with respect to the annual amount of translation,
- the corresponding terminology is available in the terminological data base (Phénix).

2 Some considerations on real texts

2.1 Notion of typology

Right from the beginning, the French-English CAT system has been developed for industrial purposes. Consequently, we have chosen to adopt a linguistic "short-cut" strategy. The baseline of this short-cut strategy is the notion of *typology.*

The French-English system was developed and tested using a chosen corpus, composed of texts from aeronautical manuals. It soon became obvious that the corpus contained a restricted number of linguistic phenomena. For instance, the corpus did not contain any interrogative sentences, pronouns (particularly relative pronouns) were very rare, etc... On the other hand, some strange phenomena regarding linguistic behaviour were discovered. For instance, certain noun juxtapositions usually forbidden in French (e.g. "l'interrupteur batterie" vs "l'interrupteur de la batterie"). Other characteristics of the texts were noted: lack of punctuation, use of many brackets...

The French-English CAT system was designed to treat all the linguistic and non-linguistic phenomena, corresponding to a particular typology of texts.

One important issue, in an industrial system, is robustness. In machine translation, robustness means robustness in the event of unforeseen phenomena. ARIANE always provides a result, even in case of failure during analysis. This is a considerable advantage, as the linguists are not obliged to take into account non-grammatical phenomena to ensure that the system will produce a translation. Some limited tests were performed on technical texts from other sources, to test the robustness of the system when faced with texts corresponding to different typologies. One important result of those tests was that many technical texts correspond to the initial typology (with minor changes from time to time).

2.2 Format of the texts

The complete integration of the CAT system in the documentation process requires taking into account all the characteristics of the texts to be handled, even the non-linguistic ones. In particular, the format of the texts should be considered.

The format of the text consists usually of non textual information, which is not always relevant from a linguistic point of view. This information is generally indicated in the document by some textual marks which indicate the logical organisation of the document. Let us quote SGML (Standard Generalized Markup Language) which is gradually becoming the standard as regards external formatting of texts.

Taking the translation problem into consideration, most of the information contained in the markup language is not relevant. Only certain types of markup indicating titles or enumeration carry linguistic information. To facilitate the translation process, it is thus important to deal with all these markup signs before calling up the CAT system.

An automatic pre-processor has thus been designed, to deal with the three main non-purely linguistic kinds of information contained in the real documentation:

- automatic conversion of the text markup (titles...),
- automatic handling of the revision marks,
- extraction of plates (tables, schemata...).

The conversion of the text markup will provide a very simplified text format, adapted to the CAT system's needs. In particular, this format will only keep in the text the markup which carries a linguistic interest (i.e. which can be used by the CAT system), other markup signs will only be stored in order to be re-inserted after translation.

The revision marks indicate which parts of the document have been changed since the last translation. In this case, only these parts of the text will be translated again.

3 The test bench

3.1 Principle

As we said previously, the 6-test consists of translating technical manuals (5000 pages) in the aeronautical field. There will be 3 kinds of manuals :

- *manuel pilote :* for the pilot's use,
- *manuel d'utilisation :* describing the aircraft parts and their functioning,
- *manuels et procédures d'entretien :* working sheets, explaining how to maintain and repair the aircraft.

The revisers will be translators working in the SITE translators' team which comprises about 60 people. The CAT system is installed on a mainframe in Paris. The translators will use a PC linked to the mainframe to call up the CAT system, through an ergonomie user interface handling all the connection problems automatically .

3.2 Users' training

First we plan a short training period (3 days) for the translators who are going to revise the rough translation. This training will comprise some general information about the machine translation system and practical information on the use of the different tools.
The training will be organised into four parts :

- introduction to the CAT system :
  how it works, what dictionaries and grammars are...
- use of the tools :
  how to call the CAT system, when and how to retrieve the translation...
- reading of the raw translation :
  how to interpret the special signs (such as double brackets), what quality of translation can be expected...
- use of the problem sheets :
  how to fill in the sheets...

3.3 Bringing into play of the ß-test

Three kinds of people will be involved in the fi-test organisation : the translators, to revise the rough translations and report back their comments to the linguist or the lexicographer, the linguist, to take care of the translators' comments, and the lexicographer, to build up the dictionary, if necessary.

The scenario of the ß-test can be described as follows :

1- human post-editing of the translation :

(a) reading of the raw translation and filling in of the problem sheets,

(b) preparation of a terminological sheet for the lexicographer, in case of incorrect translation of a terminological term,

(c) the sheets are sent to the linguists (developers) and to the lexicographer, respectively;

2- handling of the sheets by the linguists :

(a) study of the problem,

(b) correction, if possible, and corresponding elementary tests,

(c) then the sheets are sent back to the reviser(s);

3- handling of the sheet by the lexicographer :

(a) verification of the correct translation,

(b) adding of the term to the lexical data base.

3.4 Statistics

The evaluation of the CAT system will consist of four issues :

1. the translators' feeling, i.e. how they consider the system,
2. the system's efficiency, in terms of time, cost, etc...
3. the translation quality, in terms of grammatical coverage and stylistic quality,
4. the system's up-grading capability.

The translators' feeling will of course be a subjective notion, difficult to base argument on. However we do not want to neglect this aspect of the problem, as we consider that it is one of the most important aspects: a system will never be efficient if nobody wants to use it!

The second point will be easier to evaluate, as it is based on calculable figures. To prepare such calculations, it will be necessary to record some statistics automatically during the ß-test :

- the time needed by the CAT system to translate one page,
- the time needed to post-edit one page,
- the time needed by the linguists to treat all the problems corresponding to one page,
- the processing time needed for the tests and corrections (execution, compiling...),
- the servicing time: elapsed time between the date the problem was reported (by the reviser) and the updating of the system.

The third point will be obviously the most difficult to evaluate, especially as far as style is concerned. On the other hand, the grammatical coverage would be better evaluated using a test suite and tests of this kind have already been carried out by the linguists for grammar debugging purposes. In fact, the grammatical coverage corresponds to the typology accepted (as defined previously).

Last, but not the least, the up-grading capability will be evaluated in a very simple way:

- firstly, if up-grading is correct, we should receive fewer and fewer problem sheets sent by the revisers,
- secondly, we plan to re-translate some parts of the document, at the end of the test bench.

3.5 Problem sheet contents

The problem sheets comprise 4 parts:

- the sheet identification,
- the type of problem,
- the reviser’s comment(s),
- the linguist’s reply.

The first three are filled in by the reviser and the last by the linguist. Each of these parts contains several areas to be filled.

The detail of the sheet content is given in the annex.

3.5.1 Sheet identification

Three types of data are required for the identification of each problem sheet. **Initiales du réviseur (reviser's initials), Nom du texte (Name of the text)**, and **Numéro d'ordre de la fiche (Sheet sequence number)**.

The **Nom du texte (Name of the text)** is the name of the file containing the text (in French).

The **Numéro d'ordre (Sheet sequence number)** is a sequence number initialized at 1 for each text by each reviser. Any problems encountered in a given text should be indicated in order of appearance within the text. This will enable easier location within the text when reading a problem sheet. If after having filled in a sheet for two problems bearing contiguous sequential numbers (2 and 3, for example), it is necessary to indicate the presence of another problem between them, another sheet is used bearing the smaller of the two numbers to which the letter A will have been added (2A in our example). The entire alphabet can be used for the indication of problems of this nature (2B appears before 3 but follows 2A).

Other information required in this section is the **Date** of the drawing up of the sheet.

3.5.2 Type of problem

In this section, it is necessary to class the type of problem encountered into one of the categories provided. In so doing statistical calculations can be made from the information obtained. The different categories are the following:

- incorrect translation of word or expression:
  when the translation of a word is incorrect, or when an expression carrying a particular meaning and requiring a specific translation is incorrectly translated.

- incorrect word order:
  when the words of a translated sentence are not in their correct order.

- addition or suppression of words:
  when words in the translation seem to wrongly and randomly appear or disappear (in relation to the sentence or phrase in French) .

- incorrect syntax in English:
  the syntax of the translated sentence or phrase (in English) is incorrect.

- truncated word:
  for words which are translated without their corresponding suffix (i.e.

only the base form of the word appears, for example the noun ‘security’ which derives from the adjective ‘secure’ appears as ‘secur’).

- incorrect disambiguation of French:
  let us imagine that the word ‘danse’ appears in the French. The word is ambiguous as we have no way of knowing whether we are dealing with the noun or the conjugated form of the verb. The choice of one possible form of the word over the other constitutes what is termed as disambiguation.
  There is incorrect disambiguation if the wrong interpretation of an ambiguous word is given in French. The result may be the wrong translation of a given word, the explanation being that the word in English is the translation of the (wrong) interpretation obtained in French.

- absence of disambiguation in French:
  when for some reason it has not been possible to choose from the various possibilities in French. In such cases, the various possibilities are presented in English in the following form : *term1(((term2))).*

- error in the French text:
  when there is incorrect translation resulting from an error or errors in the original French text (spelling mistakes; for example the preposition à without its accent is translated as *has).*

- unknown term:
  when a word is unknown in either of the two languages (ie. it does not exist in the dictionaries) we obtain a <*term*> in the translation. The said term is either the word in French, or an internal form of the word in French or in English.

- typographical problem:
  problems in typography often occur, for example an upper case letter at the beginning of a sentence which does not appear in the translated text : “Mettre l’avion sous tension” which becomes “energise the aircraft”), or the incorrect positioning of inverted commas (eg. Mettre l’inverseur “STBY PUMP” sur “ON” which becomes ‘place’ the “STBY PUMP” switch on the ON position etc).

- other:
  for safety purposes to handle any rare phenomena which as yet have not been defined.

Following this, the reviser is asked to evaluate the sentence or part of the sentence handled in the sheet. A choice must be made from the following

- unintelligible sentence,
- intelligible sentence, but poor English,
- intelligible but clumsy sentence,
- intelligible sentence, but with a completely different meaning.

3.5.3 Reviser's comment

Translation errors are identified in this section using the following four zones:

- **Texte français (French text)**
  The part of the text in French containing the problem is written out here. Although there may be elision of certain words (indicated by "..."), location should be relatively easy. In order to be able to give a clear explanation of a problem, it may be necessary to underline or encircle words.

- **Traduction obtenue (Translation obtained)**
  This zone should contain the machine translation of the text included in the above zone, in other words, the text with or without translation errors. The same conventions are to be respected as those used for the writing of the text in French: "..." to indicate elision, underlined words, encircled words etc.

- **Traduction souhaitée (Translation desired)**
  The reviser must indicate the correct translation of the problem text.

- **Remarques (Comments)**
  It may be necessary to give a clear and precise explanation of the problem as well as a translation of the text concerned. Any comments are welcome no matter what form they may take (drawings, words etc.).

3.5.4 Linguist's reply

The above three sections of the problem sheet are to be filled in by the reviser. They appear on the front side of the sheet. The back side of a problem sheet contains any information the developer may wish to communicate to the reviser. The reviser is informed of the various measures taken in order to solve the problem or problems. The information to be given in this section appears in bold letters.

Date de réception de la fiche (Date of reception of the sheet), Date d'envoi de la réponse (Date on which the reply is sent back), Initiales du développeur (Developer's initials), for obvious reasons.

Given that several cases of erroneous translations may be indicated on a single problem sheet, the developer will define a chronological order to enable easy understanding with regard to his or her reply to the reviser. The developer rewrites the sentence using different figures to indicate the various errors in the **Numérotation d'erreurs (Error enumeration)** section.

Erreurs corrigées (Errors corrected) can have three possible answers, "yes", "no", or "partially", depending on whether all, none or only some of the errors indicated have been corrected. If the answer here is "no" or "partially", for each uncorrected error, the developer indicates why correction was not possible.

In the first instance, if the absence of correction is due to one of the following reasons (several reasons can be given):

- non-typological phenomenon,
- phenomenon which is too complex,
- low priority phenomenon,
- very rare phenomenon with undesirable side effects, or
- phenomenon which it is impossible to correct,

the number corresponding to the error will appear in front of its corresponding clause. Several numbers can appear in front of a single clause.

The absence or the wrong indexing of a term or expression is the other case where errors are not corrected by the developer. This is indicated in the following zone.

Erreur due (Error due to)

- the terminological dictionaries,
- the general dictionaries,
- the grammars.

As in the previous zone, the error numbers appear in front of their corresponding clauses. Dictionary errors can be due to the fact that a word or an expression does not exist in the dictionary, or simply that a word or expression has been wrongly entered in the dictionary (lack of or wrong information). Errors resulting from the terminological dictionaries will not be corrected by the developer. Special sheets will be prepared for such cases (build-up sheets), and will be handled by the terminological lexicographers.

An error is due to the grammars if it is not a dictionary problem. In such cases, this may be the result of a real error within the grammars, or simply untreated phenomena.

The **Remarques (Comments)** zone contains explanations concerning errors and their corresponding solutions (or the fact that no solution apparently exists). Explanations are given in this zone if the reviser is unable to obtain the desired result (indicated in **Traduction souhaitée (Translation desired)**).

Finally, the developer will indicate in the **Nouvelle phrase anglaise (New English sentence)** zone the translation obtained after modification. This may not necessarily correspond to the translation desired by the reviser.

4 Conclusion

Unfortunately, the test bench has not yet been set up. Only very short tests have been carried out (on about 100 pages). The quality of translation seems to be reasonably satisfying as far as the translators are concerned, although the terminology was not in the system at the time of testing. We are thus relatively confident regarding the translators' reaction to rough translation.

We are very fortunate to be able to collaborate so closely with the translators, and we think that this is one of the most important points which really enable the improvement of the CAT system. I would like to stress the fact that the translators are willing to participate and to use the system. In their view, it is an interesting opportunity for them and they are curious about the nature of such an MT system.

In conclusion, in my opinion this kind of test is probably possible only as an *in-house test,* as we require the revisers to spend some time filling in the problem sheets. For customers, an automatic method should be found (eg. keep all the changes made by the posteditor...).

FICHE D'INCIDENT

Initiales du réviseur :	**Date :**
Nom du texte :	**No. d'ordre :**

Type d'incident :

Cochez une ou plusieurs cases pour décrire le(s) type(s) d'incident rencontré(s)

- □ mot ou expression mal traduits
- □ ordre incorrect des mots
- □ ajout ou suppression de mots
- □ mauvaise syntaxe anglaise
- □ mauvaise désambiguïsation du français
- □ désambiguïsation du français pas faite
- □ erreur du texte français
- □ terme inconnu

Donnez votre appréciation de la traduction à l'aide des propositions suivantes

- □ phrase incompréhensible
- □ phrase compréhensible, mais mauvais anglais
- □ phrase compréhensible, mais trop lourd
- □ phrase compréhensible, mais avec un tout autre sens

Texte français :

Traduction obtenue :

Traduction souhaitée :

Remarques :

Date de réception de la fiche :

Date d'envoi de la réponse :

Initiales du développeur :

Numérotation d'erreurs :

Erreurs corrigées : □ oui □ non □ partiellement

Si non, raison pour laquelle elles n'ont pas été corrigées :

__________ phénomène hors typologie

__________ trop complexe

__________ pas prioritaire

__________ trop rare et effets de bord indésirables

__________ infaisable

Erreur due

__________ aux dictionnaires terminologiques

__________ aux dictionnaires généraux

__________ aux grammaires

Remarques :

Nouvelle phrase anglaise :

FICHE D'ENRICHISSEMENT

Initiales du réviseur :

Date :

No. de fiche :

Source	Français	Anglais	Code Objet