MACHINE TRANSLATION : WHAT TYPE OF POST-EDITING ON WHAT TYPE OF DOCUMENTS FOR WHAT TYPE OF USERS

Anne-Marie LAURIAN
Centre National de la Recherche Scientifique
Université de la Sorbonne Nouvelle - Paris III
19 rue des Bernardins, 75005 Paris (France)

ABSTRACT

Various typologies of technical and scientifical texts have already been proposed by authors involved in multilingual transfer problems. They were usually aimed at a better knowledge of the criteria for deciding if a document has to be or can be machine translated. Such a typology could also lead to a better knowledge of the typical errors occuring, and so lead to more appropriate post-editing, as well as to improvements in the system.

Raw translations being usable, as they are quite often for rapid information needs, it is important to draw the limits between a style adequate for rapid information, and an elegant, high quality style such as required for information large dissemination. Style could be given a new definition through a linguistic analysis based on machine translation, on communication situations and on the users' requirements and satisfaction.

I. MACHINE TRANSLATION AND POST-EDITING, A EUROPEAN EXAMPLE

Machine translation is often considered as a project, an experimental process, if not an impossible dream. Translation theoreticians would say no machine can understand the meaning of a text and re-express it in an other language, so no machine can translate. The debate is about the necessity of a deep semantic understanding for translating, opposed to a language structure knowledge to be sufficient to produce a translation. The usual debate is thus about the ideal concept each one has of what a translation should be.

Translation can only be defined in particular situations, regarding particular documents. And machine translation is only to be used for certain types of documents to be handled a certain way.

My observations are based on several studies I carried out on the SYSTRAN output produced in Luxembourg within the Commission of the European Communities.

In Luxembourg the amount of documents to be translated is not only very big, it is also growing very fast. The european rule is that all official documents have to be translated into the seven official languages; technical documents needed for conferences or experts meetings are sometimes translated only in three or four languages (english, french, german, italian). The delay available is often very short. That led the C.E.C. General Direction for Multilingual Transfers to promote machine translation. When they started it, some six years ago, SYSTRAN was the only system ready to produce translations. This system, originated in the U.S., has then been developed for the proper use of the Commission.

The output was far from being perfect, far from being usable as it was. Post-editing was being done. Even with the huge progress of the output quality, post-editing is still necessary. It will, in fact, be always necessary because as people get used to their translation to be done by a computer, their requirements are becoming more precise. The errors one would admit at an experimental stage, are no more possible at a productive stage.

Post-editing is thus becoming a new specialization within the numerous fields related to translation.

II - A TYPOLOGY OF DOCUMENTS BASED ON M.T. ERRORS

All documents are not suitable for machine translation. Lots of negative reactions against M.T. have been induced by a wrong use of M.T. Aware of the necessity of differentiating the documents, people responsible for translation proposed several types of typologies. They were mainly based on the subject field of the text, on its function, on its structure, on the sentence and paragraph length and complexity, on the use of particular terminologies.

The aim was to enable the chief of a translation division to choose which texts were to be sent to a human translator, and which could be processed by M.T.

My study of the errors remaining in the raw translations led me to propose a strictly linguistic typology.[1]

There are three major types of errors :

1. errors on isolated words,
2. errors on the expression of relations,
3. errors on the structure and on the information display.

These errors are classified in three tables :

1.1 vocabulary, terminology
1.2 proper names and abbreviations,
1.3 relators : - in nominal groups,
- in verbal groups,
1.4 noun determinants, verbal modificators ;
2.5 verb forms (tense),
2.6 verb forms (passive/active) and personalization (passive/non personal),
2.7 expression of modality or not,
2.8 negation ;
3.9 logical relations, phrase introducers,
3.10 words order,
3.11 general problems of incidence.

The relative frequence of these errors can be read in my tables.

These tables can be used to evaluate the probable quantity and location of errors existing after M.T., i.e. the probable quantity, location and type of post-editing. With a short training in linguistics, anyone could get trained to use these tables. By a rapid reading of the documents to be translated on the basis of these features, and according to the relative frequence of one category of probable errors or the other, one could then easily evaluate if a document should be translated by a translator or is suitable for M.T.

III - TYPES OF POST-EDITING

The system used in Luxembourg is still being developped. That means that errors are getting fewer. For instance three years ago verb forms were translated "form to form", now new rules have been introduced in order to get a past tense for a present tense (or reverse), a passive form for an impersonal one (or reverse), a.s.o.

But at the same time the variety of documents machine trabslated is growing. That means new sources on errors (mainly vocabulary, but also modalities, structures, a.s.o.).

Post-editing is always necessary. Until now post-editing has been done by translators who are wishing to do it. The amount of post-editing to do is increasing every day, it becomes obvious that post-editing can't be done just according to somebody's feeling of language and style. There has to be some rules.

Post-editing is not revision, nor correction, nor rewriting. It is a new way of considering a text, a new way of working on it, for a new aim. In order to define the characteristics of post-editing, I carried out a study on the two major types of post-editing as they appear in the C.E.C.[2]

1. The conventional post-editing (C.P.E.) is supposed to produce a text as similar as possible to what a human translation would have been, that means a high quality text.

2. The rapid post-editing (R.P.E.) is supposed to produce a correct text (on the language level as well as on the level of the meaning) but without taking care of the style.

In the experiment I carried out, time required for post-editing was the only criteria to differenciate these two methods. It appeared that special linguistical attitudes were induced by time limitation. A statistical survey of C.P.E. and R.P.E. shows the limits between :

1. necessary post-editing,
2. possible post-editing,
3. superfluous post-editing.

First group includes all post-editing that has to be done to make the text understandable, clear, readable, exact. Second group includes some research in style focused on the adaptation to the communication situation, to the author and to the presumed reader. Third group is post-editing done by people who didn't want to admit that perfection was not the aim, and that a document that will be read quickly and thrown away immediately does not require the same style as a document that will be published and largely distributed. These people usually could not give out their R.P.E. in the limited time allowed for it.

---

[1] cf. A.M. Loffler-Laurian, Pour une typologie des erreurs dans la traduction automatique, in MULTILINGUA, 2-2 (1983), 65-78

[2] cf. A.M.L.L., Post-édition conventionnelle, post-édition rapide, vers une méthodologie de la post-édition, to be published.

In rapid post-editing one has to focus on the central information, and is naturally kept out from the temptation of rewriting the sentence were errors occur. Then the post-editor finds the shortest solution, which is usually the right one. By staying very close to the raw translation, post-editors succeed in giving a good and acceptable translation.

Those who, after having post-edited according to the minimal requirements, try to make the text fit better the usual style they know, give us indications to point out the difference between :

- a text that is correct according to standard language rules,
- a text that obeys the usage rules in use on that level of documents or level of language (some "sub-rules" specific to some specialized fields, authors, situations).

IV - STYLE, SITUATIONS AND USERS

Style in literature is usually defined as the specific way an author writes. Do technical and scientific documents have a specific style ? Many people would agree on the idea that these documents have no style -or have a neutral style-. In terms of linguistic features, they can be described as well as any other writing. However the non-apparent aspect of style in informative documents is an important component of their ability to be machine translated. In a novel, the style of the author would be its main value whereas in an informative document, the transparency of style, its leaving the reader unaware of it would be essential. Even more : if style were to be felt, the information would most probably loose some of its accuracy and credibility.

In every translation situation the author has some information to transmit to a user. Let it be a technical or a political information, a scientific or a social information, the goal may be double : have the reader know more about a question (that relates to didactics), and have the reader react in a specific way to the text. Regarding this second goal, the best style, most adequate, would be the one that would bring the reader to the point the author wanted him. The neutrality of a computerized system is quite fitted to that situation. And the minimal post-editing creates often the best style.

The users' satisfaction should be the ultimate criterion to evaluate the adequacy of a style.

Are readers getting used to some new style based on machine translation ? Some people fear for the future of their language: it could evolve uncontrolled because of a new kind of users getting used to some new variety of language induced by a new tool for translation. They fear a loss of some linguistical property. Languages have always been exposed to multiple influences (wars, invasions, economical trends, cultural exchanges, a.s.o.). They are now exposed to technical influences.

Machine translation is already used by translation services. It will certainly be soon used by private translators (various systems are developped or under development in several countries). It could be used with great profit by linguists and professors to help them think about their own use of language, about the varieties of specialized uses of language, and about the future programmes that could be built up for new generations of students.

REFERENCES

- MULTILINGUA, a journal of interlanguage communication, Mouton publishers,

  see : G. Van Slype, 1-4 (1982), 221-237
  A.M. Loffler-Laurian, 2-2 (1983), 65-78
  I.M. Pigott, 2-3 (1983), 149-156

- CONTRASTES, a journal of contrastive linguistics, ADEC publisher,

  see : J. Humbley, N° 7, Nov. 1983, 35-47
  M. King, N° A3, 1983, 53-59
  A.M. Loffler-Laurian, S. Krauwer & L. Des Tombe, M.C. Bourquin-Launey, X. Huang, G. Bourquin, J.L. Vidalenc; R. Johnson, J.M. Zemb, N° A4 ("Traduction automatique - aspects européens"), 1984, 167 pp.