Comparative Evaluation of the Linguistic Output of MT Systems for Translation and Information Purposes

Elia Yuste-Rodrigo
Computerlinguistik, IFI, Universität Zürich
Winterthurerstrasse 190
CH-8057 Zürich
Schweiz – Switzerland
yuste@ifi.unizh.ch

Francine Braun-Chen
Machine Translation Management Team
Translation Service
European Commission
Luxembourg
francine.braun-chen@cec.eu.int

Abstract

This paper describes a Machine Translation (MT) evaluation experiment where emphasis is placed on the quality of output and the extent to which it is geared to different users' needs. Adopting a very specific scenario, that of a multilingual international organisation, a clear distinction is made between two user classes: translators and administrators. Whereas the first group requires MT output to be accurate and of good post-editable quality in order to produce a polished translation, the second group primarily needs informative data for carrying out other, non-linguistic tasks, and therefore uses MT more as an information-gathering and gisting tool.

During the experiment, MT output of three different systems is compared in order to establish which MT system best serves the organisation's multilingual communication and information needs. This is a comparative usability- and adequacy-oriented evaluation in that it attempts to help such organisations decide which system produces the most adequate output for certain well-defined user types.

To perform the experiment, criteria relating to both users and MT output are examined with reference to the ISLE taxonomy. The experiment comprises two evaluation phases, the first at sentence level, the second at overall text level. In both phases, evaluators make use of a 1-5 rating scale. Weighted results provide some insight into the systems' usability and adequacy for the purposes described above.

As a conclusion, it is suggested that further research should be devoted to the most critical aspect of this exercise, namely defining meaningful and useful criteria for evaluating the post-editability and informativeness of MT output.

Keywords

MT Quality, Users' needs, Administrators, Translators, Informativeness, Post-editability.

0. Introduction

The evaluation experiment described in this paper was carried out by a team of five linguists/translators last April during the ISLE[1] MT Evaluation Workshop at ISSCO[2], University of Geneva, Switzerland.

Focused on the particular needs of two well-defined user classes – translators and administrators – within a multilingual organisation (the European Commission, or EC), the experiment sought to compare the *output quality* generated by three different MT systems.

The proposed model was a *black-box* type *usability* (White 2000) evaluation. In other words, there was no interaction with the systems tested, and the goal was to determine whether output was actually helpful to the user groups in question. The term *usability* is considered here as akin to the Commission's notion of *adequacy*, which we also wish to embrace: "assessing the adequacy of a system/systems with respect to the users' requirements within the Commission's environment" (EC Translation Service 1998).

In keeping with the ISLE taxonomy, we start off with a brief description of our reference population's needs. Section 1 therefore presents the European Commission's MT users and workflow.

After reporting on the experiment itself, we then propose a reduction in the criteria applied when evaluating the linguistic output of MT in an international organisation. We also offer some guidelines on how and what to evaluate when it comes to *post-editability* and *informativeness.*

1. User Needs

1.1 User Population

Currently there are eleven official languages in the European Commission: Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.[3] With accession of the Central and East European countries to the EU, translation from and into at least ten new languages will somehow have to be provided.

The Commission's has a total of over 18,000 staff, of whom some 1,300 are translators - indication of how important translation-related activities are.

1.2. Translation Activity

[1] *International Standards for Language Engineering* - http://www.issco.unige.ch/projects/isle/mt-eval-programme.html

[2] *Institut Dalle Molle pour les Etudes Sémantiques et Cognitives* - http://www.issco.unige.ch/index.html

[3] This amounts to 110 possible language combinations. The MT system used by the EC covers 18 language pairs of varying translation quality. The quality needs of translators and administrators will often coincide, but if output from the less-developed pairs might already be suitable for administrators who just need to know the content of a document written in an language they do not understand, it is not likely to meet the requirements of translators seeking to gain time by post-editing the results.

More than one million pages are translated every year by human translators at the EC and, in 2000, about 420,000 pages were machine-translated[4]. The Translation Service accounted for 44% of the Commission's MT demand, with nearly 185,000 pages and 20,000 documents. In other words, translators are by no means the only users of MT within the organisation, since the remaining 235,000 pages were requested by non-linguists (administrators).

At the Commission, translation usually forms part of the document production chain right up until final publication. EU legislation, general and specialised reports and written communication, be it within the institution or with the outside world, represent the bulk of the translation activity. With regard to legislative texts, there is an obligation to publish in all 11 official languages.

Since MT is readily accessible to everyone within the organisation, translators are free to use its raw output as a basis for producing translations of publication quality, that is, by post-editing it. For very urgent internal texts, they can occasionally do a lighter, *rapid* post-edit.

Raw MT output can also be used by administrators:

- as an authoring tool for drafting in a language other than their mother tongue;
- for the translation of urgently needed documents (with revision by a native speaker in the department concerned);
- or, most commonly, for browsing information in a language they are not familiar with.

In this last case, users can then decide if they wish to submit their texts (or part of them) for human translation, or whether the information provided in the raw translation is sufficient.

The role of MT in the multilingual communication of the Commission and the type of activities it involves are central to this study. Indeed, we later focus on the usability of MT output, both as a post-editable material for translators and as a source of information for administrators.

2. Evaluation Framework

2.1. Evaluators

The evaluation team was composed of five linguists/translators of French and Spanish mother tongue, with heterogeneous experience in the MT field, ranging from regular MT use in their workplace/translation workflow to corporate system development or academic training.

The amount of time available for the experiment, including preliminary warm-up discussions and final presentation, was three days.

2.2. Three systems tested

It was agreed that three MT systems would be tested: one taken from the market, the other two from international organisations, namely the EC and WHO, both of which were thus adapted to their users' needs.

[4] A further 126,000 pages was requested by other institutions and public authorities in the Member States.

2.3. Language Combinations

As time was limited, the scope of the experiment had to be reduced to a minimum. It was decided to evaluate two language pairs with the same source: English-French and English-Spanish were chosen since the evaluation team comprised native speakers of those target languages.

2.4. Text Types

A small sample of English source texts was selected for machine translation and evaluation. The four documents chosen were representative of the texts usually translated at the WHO or EC: short extracts were taken from a medical document on antibiotics, from two legal texts (on antidumping and the ACP Convention), and from a more general text (answer to a parliamentary question).

The "antibiotics" and "antidumping" extracts were translated into Spanish, whereas the "ACP Convention" and "answer" were translated into French. For Spanish target, linguistic output of the three MT systems was assessed (Commission - A, WHO - B and commercial product - C), while for French target the results were only examined for two products (A and C).

3. Evaluation Process

3.1. Preparation

3.1.1. Output Characteristics

Since our evaluation procedure was based on MT users' needs and was ultimately aimed at improving their productivity, we did not venture into the technical and economic aspects of MT systems. Rather, we compared the quality of MT output (using linguistic criteria) in order to determine whether the systems did indeed satisfy the users' specific needs.

3.1.2. Criteria Selection

Once the target user groups and their professional needs were known, it was necessary to establish valid evaluation criteria for our experiment.

We therefore proceeded to consult the ISLE MT Evaluation Taxonomy (as of April 2001). Much time was spent on selecting the appropriate criteria and metrics, a lengthy discussion proving necessary to reach a consensus among the five team-members. This highlighted the eminently subjective nature of the selection exercise itself.

It was agreed at the time that it was better to choose a variety of criteria from the taxonomy to test how valid they were in practice. However, given the limited time and lack of evaluators, we later realised that this approach had been too ambitious.

3.1.3 Criteria and Metrics

Two types of criteria were selected, one on sentence level, the other on text level.

1) We started work at **sentence level** by rating the accuracy of each translated sentence on a scale of 1 to 5, taking into account **punctuation, capital letters, morphology, lexis and syntax**.

2) We then evaluated the **overall text,** using the same scale, on the basis of the following seven criteria: **coherence, comprehension/intelligibility, fidelity, readability, style, terminology and usability**.

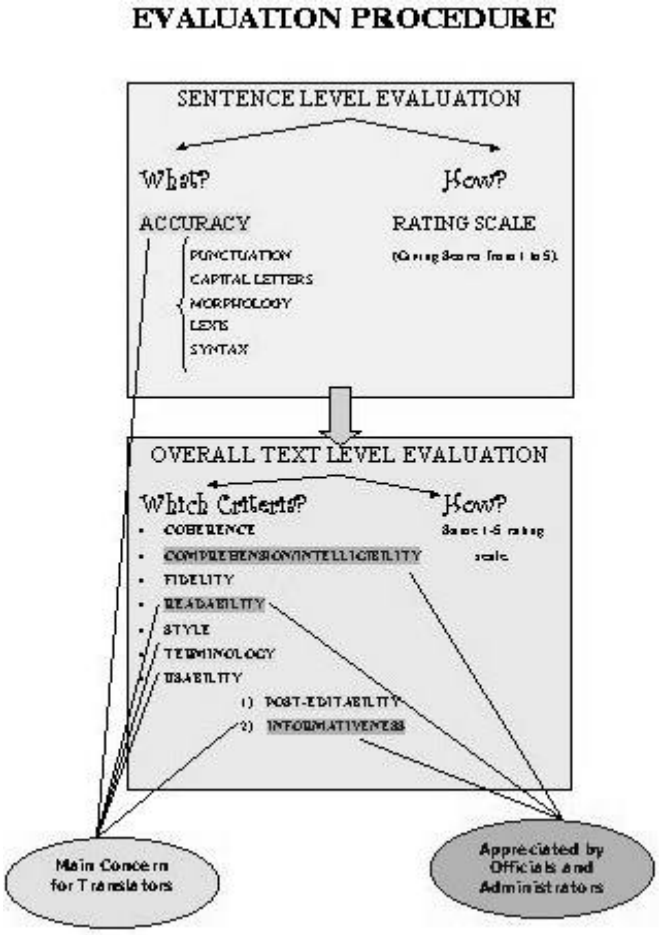

Figure 1: Evaluation Procedure: First Stages

The point of a 1-5 scale was to produce an Excel table of average scores that would help decision-makers to choose between several evaluated systems.
Figure 1 shows the two phases of this evaluation exercise. It also started to hint at the criteria that matter most for translators and administrators. This aspect will be further explained below.

3.2. Evaluation Results

3.2.1. Accuracy/Comparison at Sentence Level

An assessment of accuracy at sentence level of the raw Spanish output of systems A, B, and C suggested that A gave the best results for "Antidumping" and B for "Antibiotics". Evaluation of the two English extracts translated into French by systems A and C showed that A performed better than C since the former had been adapted to the text types concerned.

3.2.2. Comparison at Text Level

The sentence-level trends were reproduced at text level when coherence, comprehension/intelligibility, fidelity, readability, style, terminology and usability were compared: system A performed best with the EC documents, system B with the WHO text.
System C, which had not been developed specifically for any of the texts, performed least well, be it for Spanish or French target.

3.3. Observation

The evaluation experiment corroborated the hypothesis that those systems already tailored to an organisation's needs usually yielded better quality outputs than a commercial system.
As regards to criteria usage, it was felt at the time that a combination of the above criteria was important for producing a comprehensive evaluation result. It was later observed, however, that although the criteria were all fairly important, such a broad evaluation would be unrealistic and would come at a cost. So if an international organisation such as the Commission were in fact to conduct an evaluation of MT output, it would surely concentrate on fewer criteria, as each deserves a detailed experiment in its own right.

4. Main Criteria for Evaluating Linguistic MT Output – Rethinking the Evaluation Exercise

In this section we focus on a number of basic criteria for evaluating the linguistic output of MT systems, namely *accuracy*, *comprehension/intelligibility*, *fidelity*, *terminology* and *usability* (comprising *informativeness* and *post-editability*. An evaluation might incorporate any or all of these factors.

4.1. Criteria Properties

Four possible properties were assigned to the evaluation criteria: **input-dependent** (**ID** – source language must be taken into account when considering criterion), **output-dependent** (**OD** – target language must be taken into account when considering criterion), **key-for-information-purposes** (**K4IN** – concerns MT output as an information source, i.e. the results do not have to be of publication quality), and **key-for-translation-purposes** (**K4TR** – MT output regarded as an aid for producing translations of publication quality).
Table 1 indicates the properties assigned to each criterion.

EVALUATION CRITERIA PROPERTIES				
Criteria list	**ID**	**OD**	**K4 IN**	**K4 TR**
Sentence level				
ACCURACY	√	√		√
Text level				
COMPREHENSION or INTELLIGIBILITY		√	√	
FIDELITY	√			√
TERMINOLOGY	√	√	√	√
USABILITY:				
- INFORMATIVENESS		√	√	
- POST-EDITABILITY	√	√		√
ID: Input-dependent **OD: Output-dependent** **K4IN: Key for Information Purposes** **K4TR: Key for Translation Purposes**				

Table 1: Evaluation Criteria Properties

4.2. Accuracy

In the ISLE Taxonomy, *accuracy* refers to those "attributes of software that bear on the provision of right or agreed results or effects" (based on ISO 9126: 1991, A.2.1.2). When accuracy concerns the whole text, it comprises the subcategories of *fidelity, comprehension/comprehensibility, consistency, and coherence*. When applicable to sentence-level evaluations, it comprises morphology and syntax. *Accuracy* also includes a third evaluative category - a typology of errors which identifies four classes of linguistic fault: punctuation, lexis, syntax and style.
We adopted *accuracy* as the logical criterion to use for a sentence-level evaluation and proceeded to assess each sentence in terms of punctuation, capital letters, morphology, lexis and syntax.
Throughout the experiment, we noted that *accuracy* was both an input- and output-dependent criterion. We also agreed that it was *key for translation purposes*, since greater accuracy meant fewer corrections and a bigger time gain for translators.

4.3. Comprehension/Intelligibility

We consider these terms to cover broadly the same area: the ease with which MT output can be understood by the user. This is a subjective criterion which is usually *output-dependent* and *key for information purposes*, as it is precisely the ease with which administrators can get the gist of a foreign-language text that matters most. In our experiment, *comprehension/intelligibility* was generally applied to the whole text, although we had prior knowledge of specific "trouble spots" thanks to the sentence-level analysis.

4.4. Fidelity

Fidelity (the accurateness and completeness of the information conveyed, as defined in ISLE), was found to be heavily dependent on input. As Van Slype puts it, a fidelity-driven evaluation would subjectively examine "the degree to which the information contained in the original text has been reproduced without distortion in the translation." This is of immense importance to translators, who would seize on any type of information anomaly produced in the translation, i.e. *non-translated item* (loss of information or silence), *added item* (interference or noise), or a combination of the two (information distortion, such as mistranslated terms). We agree with the ISLE argument that "detailed analysis of the fidelity of a translation is very difficult to carry out, since each sentence conveys not a single item of information or a series of elementary items of information, but rather a portion of message or a series of complex messages whose relative importance in the sentence is not easy to appreciate." Nonetheless, translators/linguists performing a fidelity-driven evaluation could take advantage of automatic alignment tools in order to examine input and output alongside each other. It would still be a subjective and somewhat laborious evaluation, but automation could lessen the load.

4.5. Terminology

Terminology is said to involve a "subjective evaluation of the degree to how correctly the most important terms are translated" (ISLE Taxonomy). Here the intertextual[5] level of the criterion is underlined, that is, how terms are translated across texts from one language to another. As suggested in ISLE, *terminology* could be measured by determining "the percentage of names or other input/output domain terms that are mistranslated".
We consider that *terminology* is a highly relevant criterion for both information and translation purposes. Whilst administrators do of course benefit from a good handling of subject matter terminology by the system, it is usually translators who are in charge of researching terminology in the organisation's domain(s), compiling specialised databases and updating the MT system's dictionaries. In this study, *terminology* is therefore equally important for information *and* translation purposes. By the same token, it is classified as both *input-* and *output-dependent*, since the relevant terms must be properly described and used in all languages in which the organisation produces documents. Consequently, this is the most important criterion for evaluating MT linguistic quality in our chosen test environment, since it is of interest to every user and is endowed with all four criterion properties: ID, OD, K4IN and K4TR.

4.6. Usability (Post-editability and Informativeness)

[5] In the terminology field, it is often useful to make a distinction between **intertextual** and **intratextual** terminology levels. Whereas the first level refers to how terms are translated (i.e. term handling in one or more language pairs) and is thus input-dependent, the second refers to how terms are treated within a single text, highlighting the importance of internal terminological consistency. Although both terminology levels are essential for translation and information purposes, administrators with no knowledge of the source language necessarily depend on adequate and consistent terms *throughout* their MT text; translators are not so reliant.

Finally, our prime purpose was of course to determine the *usability* (or *utility*) of MT systems based on their *informativeness* and *post-editability (revisability)*: in a nutshell, which system was most useful for administrators and translators? The former use it for information-gathering in a foreign language (key: *informativeness*), the latter take it as a raw material for producing a polished, publication-quality translation (key: *post-editability*).

A couple of thorny questions emerge: how can informativeness of MT output be measured? And to what extent is a machine-translated text post-editable? In other words, how do you determine the cut-off point beyond which post-editing MT takes more time than translating from scratch? Shortage of time prevented us from providing any answers: that task should be the next stage in the exercise.

4.6.1. Informativeness

The ISLE Taxonomy introduces the notion of *informativeness* as "semantic fidelity", questioning whether the output reflects the content of the source text and whether distortions of meaning occur. *Informativeness* is associated with users (here, administrators) who usually do not have enough knowledge of the source language (input) to be able to make use of the source data. We therefore think that it is more appropriate to consider *informativeness* as primarily output-dependent – OD. In that sense, it might be seen as being more akin to *comprehension/intelligibility* rather than to other criteria such as *fidelity*.

4.6.2. Post-editability

Post-editability (also called *revisibility*) has been defined as "the stage at which the translated text needs to be transferred into a form, which meets the requirements of the final publications and/or delivery process" (OVUM report, quoted in ISLE). Previous studies have shown that *post-editability* is "the phase in the computer-assisted translation process that takes the most time" (Trial of the Weidner Computer-Assisted Translation System, p.12, October, 1985, quoted in ISLE). This is why post-editability is usually associated with time or speed as an evaluation parameter, the typical experiment being a comparison of MT post-editing time against human translation time: if post-editing takes longer than translating from scratch, then the MT output is considered unusable for translation purposes. Other experiments are based on the number of individual editing steps required to bring a text up to an acceptable level - classification and counting of word replacements, deleted words, sentence rearrangement, etc. - adding up to create the so-called *edit distance*. This kind of evaluation is more oriented towards examining system quality for a given text type. In fact, no matter which evaluation scenario for *post-editability* is used, this criterion will be highly dependent not only on the purpose of the output, but also on the nature and quality of the input text test. For this reason, it is claimed that post-editability is both ID and OD (see **table 1** above).

Moreover, post-editing should not be seen as an isolated task but an activity immersed in a wider and more varied cycle of machine-aided multilingual document production in large organisations, where other important activities take place (i.e. pre-editing, human-machine interaction, translation memories, etc.). These are all "strategies for optimising the quality of MT output … that do not exclude each other; they can be applied in addition to each other" (Austermühl 2001: 162-3). For instance, the post-editing stage will take less time and effort if the MT system's dictionary is constantly fed with new terms, or if texts sent to the system are pre-edited or written in a controlled language. Those performing a *post-editability*-driven evaluation should be aware of these interconnections.

Although our usability evaluation requires more development, the following correlation can already be established: the more *informative* and/or *intelligible* the MT output, the more usable it is for information purposes; the less post-editing is needed, the more suitable the MT output is for translation purposes.

5. Conclusion

Further research should be conducted on what proved to be the most critical aspect of this exercise, namely defining meaningful and useful criteria for evaluating *post-editability* and *informativeness* of MT output. Time permitting, the criteria and metrics employed here should be revisited and explored more thoroughly. The evaluation experiment did however enable us to better understand the circumstances by which the MT quality and accuracy needed for translation purposes is relatively higher than for information purposes. Besides, two general conclusions can be drawn: 1) any evaluation of MT linguistic output, even if mathematically measurable, involves a subjective factor; and 2) MT is most suitable when its development has been targeted to the needs of a specific user category.

References

ALPAC (1966). Languages and machines: Computers in translation and linguistics, National Research Council Publication 1416, National Research Council, Washington, DC. Report of the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences.

Austermühl, F. and A. Kortenbruck (2001). A translator's sword of Damocles? An introduction to machine translation. In Austermühl, F. (2001), Electronic Tools for Translators (pp. 153—176). Translation Practices Explained series. Manchester, UK & Northampton, MA: St. Jerome Publishing.

EAGLES Evaluation Group (1999). The 7-step recipe.

European Commission Translation Service (1998). *Suggestions for the Evaluation of MT Systems*. (internal document). Directorate for General and Language Matters, Development of Multilingual Tools, Luxembourg.

ISLE [International Standards for Language Engineering] (2001). *Evaluation of Machine Translation*. http://issco-www.unige.ch/staff/andrei/islemteval2/

Reeder, F. (2001). An introduction to MT Evaluation. http://www.issco.unige.ch/projects/isle/mte-introduction-fr/index.htm

Van Slype, G. (1979). *Revue critique des méthodes d'évaluation de la qualité de la traduction automatique*. *Rapport final*. Pour la Commission des Communautés européennes (DG XIII). Bureau Marcel van Dijk.

Acknowledgements

We wish to express our gratitude to the organisers of *MT Evaluation Workshop: An invitation to get your hands dirty!* (University of Geneva, 19 - 24 April 2001) for all their efforts. We should also like to thank the other workshop participants and the rest of our evaluation team - Marijo Astre, Anna Civil, and Marianne Starlander - for a fruitful exchange of ideas. We also owe special thanks to two Commission colleagues, Rosemarie Sauer-Stipperger for her continuous interest in this paper, and Cameron Ross for his invaluable linguistic proofreading. We, of course, assume full responsibility for all deficiencies, inadequacies and omissions.