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

Quality Criteria for MT

Lorna Balkan
Department of Language and Linguistics
University of Essex

Introduction

Quality criteria are used as the basis for declarative evaluation of MT, yet despite the considerable literature on the subject (Pierce et al. 1966, Nagao et al. 1985, Dyson, M. et al. 1987) there is little discussion about what constitutes a proper set of quality criteria for MT, and on what basis it is motivated. This paper aims to address these questions. In particular it asks whether this set is the same as the set standardly used for human translation, and if not, why not. The human translation quality evaluation criteria we examine are those used by the Institute of Linguists (undated) (henceforth IL), a professional examination body, and those used by the Translation Bureau of the Secretary of State of Canada (1990) (henceforth TB), which is a professional translation service. The paper is in the following parts:

1. Some quality assessment criteria for human translation
1.1. The Institute of Linguists
1.2. The Translation Bureau of the Secretary of State of Canada
1.3. Discussion

2. Some traditional quality assessment criteria for MT
3. A Comparison of Human and MT Quality criteria
4. Conclusion

1 Some quality assessment criteria for human translation.

1.1 The Institute of Linguists.

The Institute of Linguists sets an examination for its Diploma in Translation, which is intended to "assess and reward professional competence in translating from another language into English or from English into another language”. Professional competence is taken to mean that the work would be “of a standard acceptable for submission to a commercial client”. There are three classes of criterion:

1. Commenting on the translation difficulties of the source text.
   This will not concern us here, since it has no bearing on translation quality.
2. Decoding the source text
   The only criterion that falls under this heading is accuracy.
3. Encoding the target language version.
   Three criteria are distinguished here:

- choice of vocabulary, idiom, terminology and register
- cohesion, coherence, organisation
- grammar, punctuation, spelling and transfer of dates, names, figures, etc.

There are thus 5 criteria in all, each of which is marked on a 5 point scale. An aggregate grade 3 or above in each paper is required in order to pass. Evaluation is performed on the translation as a whole and not sentence by sentence.

We shall look below at the decoding and encoding criteria in greater detail.

1.1.1 Decoding.

By accuracy is meant the “correct transfer of information”. It is assumed that this requires an excellent understanding of the subject matter. Marks are deducted for errors and omissions, which are classified as either “major” or “minor” depending on whether they are “not such as to give directly false information to the reader” (minor) or lead to “information being conveyed wrongly at several points” (major). Marks are also deducted for “lack of clarity” and “lack of economy in the translation”.

1.1.2 Encoding.

A. *Choice of vocabulary, idiom, terminology and register.*

The concern here is that the language and register is “entirely appropriate to the subject matter and to the spirit and intention of the original”. Serious infelicities that “impair or distort the message” are distinguished from less serious ones which do “not impair the overall acceptability of the translation”.

B. *Cohesion, coherence, organisation.*

This criterion is concerned that “the sentence structure, linkages and discourse organisation are all entirely appropriate to the target language”. Marks are deducted for structural features that are taken over from the source text but which are inappropriate to the target language. The effect of poor structure may be that the translation sounds stilted or that there may be “some incoherence which was not present in the original”.

C. *Grammar, punctuation, spelling and transfer of dates, names, figures, etc.*

This criterion ensures that punctuation, spelling and grammar comply with the conventions of the target language as well as the representation of dates, names, and numbers. Again, a distinction is made between minor and major errors, the latter of which affect the acceptability of the translation.

Thus it can be seen that the overriding concerns embodied in the above criteria are that the translation preserves the information and register of the original, that it reads like a piece of target language text, and that it observes the linguistic conventions of the target text.

1.2 The Translation Bureau of the Secretary of State of Canada.

We now turn to consider the quality assessment criteria used by the Translation Bureau of the Secretary of State of Canada. First of all, a distinction is made between “translation” errors and “language” errors. “Translation” errors occur if the meaning of the source text has not been faithfully rendered, while language errors arise if the idea has not been formulated correctly in the target language.

All error types are also classified into “major” or “minor” errors. The severity of an error can depend upon the context (ie. whether the element in which the error occurs is an essential element or merely a secondary element of meaning in the source text), on the potential impact on the reader, or even on the nature of the text (a document intended for publication obviously requires a higher standard of quality than an internal memorandum). Thus it can be seen that the decision as to whether an error is major or minor is highly subjective and text-specific. While the reader is provided with many examples to help him distinguish between translation/ language errors on the one hand and major/minor errors on the other hand it is acknowledged that the distinctions are not always very clear.

Evaluation is performed on one or more 400-word samples of translation, depending on the length of the text. A translation is considered fully acceptable if it contains no major errors and 12 or fewer minor errors. It is considered revisable (ie. submittable to an in-house reviser) if it contains one major error and 18 or less minor errors.

We shall now look at the composition of translation and language errors in greater detail.

1.2.1 Translation errors.

Translation errors, it will be recalled, are concerned with the correct rendering of the message of the source text, i.e. with accuracy. The factors which affect accuracy include mistranslation. A minor mistranslation is one that changes the meaning of the original only slightly, while a major mistranslation is one that "drastically alters the meaning of an essential element of the message". A mistranslation is not considered major if the context allows the reader to correct the error mentally.

Accuracy can likewise be affected by omissions, as we saw above, and conversely, by additions, defined as "unjustified introduction of an element of meaning in the translation that does not appear in the source text". In addition, the introduction of ambiguity in the target text where there was none in the source text is considered a translation error.

1.2.2 Language errors.

Language errors are subdivided as follows:

- Diction ("The translator's choice of words")
- Punctuation
- Syntax
- Style ("The manner in which ideas are expressed")

and some non-categorised errors including faulty linkage between clauses and sentences.

Diction covers words or phrases that are inappropriate, either because they have inappropriate connotations or register or constitute an error in collocations or idiomatic expressions, or because they are too general or too specific (but where meaning or nuance is lost, it is considered to be a translation error).

Punctuation is self-explanatory, but a distinction is made between punctuation errors which affect the meaning of the text (which count as translation errors) and those which do not (and are therefore language errors).

Syntax is concerned with grammaticality.

Style covers errors such as unnecessary repetition of a word or idea and a too literal translation of the source text, resulting in a translation that is “unidiomatic or difficult to understand”.

1.3 Discussion.

There is considerable overlap between the criteria used for quality control by the the Institute of Linguists and those used by the Translation Bureau of the Secretary of State of Canada. The decoding/encoding distinction in the former is mirrored to a large extent by the translation/language error distinction in the latter. Both decoding and translation errors are concerned with the correct transfer of information, while encoding and language errors are concerned with how this information is expressed in the target language. Notice that viewing the translation process as involving decoding and encoding the source message is a notion which is familiar in translation theory (see for example Nida 1964).

Generalising somewhat, we can view decoding as being concerned chiefly with *content* and encoding as concerned chiefly with *form.* Following modern translation theorists (e.g. Snell-Hornby (1988), House (1981), Newmark (1981)) we take “content” to include pragmatic as well as semantic meaning. This means that the translation must convey the intention as well as the denotational meaning of the original.

The factors considered to be relevant to encoding (form) in the criteria of both the Institute of Linguists and the Translation Bureau of the Secretary of State of Canada are:

- grammar and punctuation
- idiomatic and collocational usage
- sentence structure and discourse structure
- register
- word choice

We propose to classify grammar and punctuation together as “grammaticality”, and all other aspects of encoding as “style”, which we define as follows:

“The conventional patterns of expression which characterise particular languages”.

While some factors of style (e.g. correctness of idioms and collocational expressions) are common to all types of language use, other factors of style (e.g. word choice and sentence structure) are dependent on the particular type of sublanguage being used, where sublanguages vary according to:

- subject matter (e.g. business)
- field of discourse
  (this reflects the social function of language, e.g. a letter)
- mode of discourse
  (broadly, this is the distinction between written and spoken language)
- tenor of discourse
  (this reflects the degree of formality of language and is what appears to be meant by “register” in the criteria of the IL and TB).

There is another aspect to style, namely the linguistic habits of an individual user, be they conscious (eg. the deliberate use of repetition as a kind of leitmotiv) or unconscious (eg. geographical or social variation). We propose to distinguish these aspects of style which we shall call “literary style” from the aforementioned aspects of style which we shall call “conventional style”, since only in literary texts is it necessary to render “literary style” in translation.

We have been assuming up until now that we can separate form and content. Yet this is not quite true, as translation theorists acknowledge (e.g. Nida 1964, p154): “Of course, the content of a message can never be completely abstracted from the form, and form is nothing apart from content”. Some uses of register, for example, convey information about the social relationship between the addresser and the addressee. Newmark (1988, p14) illustrates the difference between formal and informal registers with the following examples:

Formal: “You are requested not to consume food in this establishment”.
Informal: “Please don’t eat here”.

Nevertheless, we shall continue to treat all the aspects of style we listed above as matters of form, since with the exception of “literary style”, most are a matter of convention, and where they are not, as with register, the meaning difference is not such as to radically alter the meaning of the source text.

That there is a link between form and content is recognised in the criteria of the IL and TB. Thus the IL criteria associate cohesion (a syntactic term) with coherence (a semantic term); a translation may be “stilted and incoherent”.

It is useful to think of decoding as an essentially bilingual operation, which requires reference to both the source and target text, and to think of encoding as principally a monolingual criterion that requires reference to the target text only. Again, this view requires qualification. Some aspects of style do require reference to the source text. Thus in IL, register must be “faithful to the register of the source text”, and in TB “words or expressions that are inappropriate to the style of the text. They may be either excessively colloquial or excessively formal.”

Summarising, examination of the IL and TB criteria provide us with a way of viewing the translation process, namely in terms of decoding and encoding a text. From this we derive three criteria for assessing translation quality, the first of which is associated with decoding, and the remainder of which are associated with encoding:

1. accuracy
2. style (conventional and literary)
3. grammaticality.

2 Some traditional quality assessment criteria for MT.

We now look at some quality criteria that are standardly applied to MT. Perhaps the best known example of quality evaluation for MT is J.B. Carroll’s contribution to the US Government ALPAC report (see Pierce et al. 1966). Carroll uses two criteria, accuracy and intelligibility, both of which are measured on 10-point scales. No definition of either is provided, but it is clear that Carroll has in mind two distinct criteria: “a translation could be highly intelligible and yet lacking in fidelity or accuracy. Conversely, a translation could be highly accurate and yet lacking in intelligibility.” (Carroll op. cit. p67).

What, then, does intelligibility consist of? On the one hand it talks about “comprehension”, on the other hand it talks about style and grammar, but all terms are used without reference to the source text. Intelligibility would appear therefore to belong to the encoding process rather than the decoding process, since we saw that encoding criteria can be assessed in principle on the basis of the target text alone. However, while the encoding criteria we looked at for human translation, namely style and grammaticality, were chiefly concerned with form, Carroll's intelligibility scales talk about comprehension, a content-related word. There are in fact two aspects to encoding: WHAT is encoded (content), and HOW it is encoded (form). While the human translation encoding criteria discussed above (style and grammar) stress *form* Carroll's intelligibility scales appear to stress *content.* The mistake Carroll makes is that he equates comprehension (content) with style and grammar (form). Thus for a text to be considered intelligible it must be "perfectly clear" and have no "grammatical or stylistic infelicities". As we showed above, there is a link between form and content, but the link is not as straightforward as Carroll's scales suggest. Lack of comprehension cannot simply be put down to poor style or grammar. Thus poor style may interfere with what is being expressed, but this is not always so, e.g.

he read the guide of the user
= he read the user guide

Likewise, poor grammar may render a sentence incomprehensible, but this need not be the case, e.g.

she has said me that it has two houses
= she has said to me that it has two houses

In fact, lack of comprehension may be also be totally unrelated to either style or grammar, e.g.

I like to swim by tomorrow

What we need to do is to define intelligibility independently of style and grammar. We must nevertheless acknowledge that both style and grammar MAY affect intelligibility. Intelligibility on this account would be purely content-related and mean something like "comprehensibility". Van Slype (1982), who criticised Carroll for including style in his intelligibility scales, came up with his own intelligibility scales which seem implicitly to embody the idea that intelligibility is independent of both style and grammar. Thus for a text to be "very intelligible" "all the content must be comprehensible, EVEN THOUGH there are errors in style and/or spelling.." (op. cit. p.226) (translation and emphases are mine). We would further suggest that intelligibility can only be assessed on the basis of an entire text (or section of text), and not on random sentences, such as were presented to the evaluators in Carroll's test. This is because under normal circumstances much of our comprehension of language is derived from context.

Notice that the definition of intelligibility that we are adopting assumes, like Carroll, that intelligibility is a monolingual criterion (i.e. that it can be assessed without reference to the source text). Other researchers take the view that intelligibility is a bilingual criterion. For example Sager (1989, p97): “This global approach includes such criteria as intelligibility, which must be measured against the intelligibility of the original..” Of course, it is undeniable that an unintelligible source text may produce an unintelligible translation. If we want to retain intelligibility as a monolingual criterion, then we must make the assumption that the source texts are intelligible to a high degree. Any lack of intelligibility will therefore be attributable to the translator or the translation machine. We thus keep intelligibility as a monolingual criterion, since we think it is useful to have some criterion for assessing the content of the target text independently of the source text. Thus for MT we have basically two criteria that are standardly used for MT:

1. accuracy
2. intelligibility

with possibly two further relevant criteria:

3. style?
4. grammar?

3 A comparison of human and MT quality criteria.

While for human translation we found three standardly used criteria, namely accuracy, grammaticality and style, for MT we found two, namely accuracy and intelligibility. Below we consider if style and grammaticality are relevant to MT, and conversely, whether intelligibility is of relevance to human translation.

First of all we need some way of measuring “relevance”. Translations do not exist in a vacuum, but are produced for a purpose. An assessment of the relevance of quality criteria must take due consideration of this purpose. Sager (1989, p91): “There are no absolute standards of translation quality but only more or less appropriate translations for the purpose for which they are intended”. It so happens that the purpose to which a human translation is put is often very different to that to which a machine translation is put. Human translation, for example, must usually be of a sufficient quality to be “of a standard acceptable for submission to a commercial client” (Institute of Linguists p.l). This normally means that not only must the translation be accurate, but it must also be presented in good, correct English. This lays the emphasis on style and grammar. It will normally be assumed that the human will produce intelligible output, so it is not considered necessary to evaluate intelligibility separately from style and grammar.

In MT, however, it is usually the case that raw MT output is not presented directly to the client but is first post-edited. The post-editor may correct any stylistic and/or grammatical inaccuracies (see Laurian 1984 for a discussion of some different levels of post-editing). Thus, the overriding criteria for MT must be accuracy and intelligibility. This is particularly the case where the raw MT output is presented to the client directly for the purposes of "gisting". Note that an assessment of intelligibility would normally precede an assessment of accuracy, since if something is not intelligible, it is impossible to say whether it is accurate or not.

This is not to say that the assessment of style and grammaticality have no place in MT evaluation. On the contrary, while in the past few MT systems made claims about producing grammatical or stylistic output, more modern systems which perform full syntactic analyses of sentences (such as EUROTRA, see Arnold et al. 1987) can be expected to produce grammatical output. Thus some criteria must be available to assess this, namely the grammaticality criterion.

There are even attempts nowadays to investigate how MT systems can be made more sensitive to stylistic conventions (see Dimarco 1990). Again, style must be retained as a criterion if this is to be judged. However, it is probably fair to say that the assessment of "literary style" (as opposed to "conventional style", as defined in section 1 above) will never become relevant to MT, as MT is neither intended for, nor suited to, the translation of literary texts.

4 Conclusion.

Examination of standard human translation quality criteria suggested two aspects of translation, namely "decoding" and "encoding" from which quality criteria could be derived, namely "accuracy" "style" and "grammaticality". Traditional MT evaluation on the other hand has mostly centred on "accuracy" and "intelligibility". We attempted a definition of all four criteria and considered their relevance to both human translation and MT by considering the context in which each type of translation is used. We concluded that intelligibility is of marginal importance in human translation, while all four criteria (with the exception of literary style) are relevant to MT.

REFERENCES

Arnold, D. and des Tombe L. (1987): “Basic theory and methodology in EUROTRA” in Nirenburg, S. *Machine Translation: Theoretical and Methodological Issues,* CUP, Cambridge.

Dimarco, C. and Hirst, G. (1990): “Accounting for Style in Machine Translation” in *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language,* University of Texas, Austin, June 1990.

Dyson, M. and Hannah, J. (1987): “Towards a Methodology for the Evaluation of Machine-Assisted Translation Systems”, in *Computers and Translation,* Vol. 2, no 3, pp163-176.

House, J. (1981): *A Model for Translation Quality,* Gunter Narr Verlag, Tübingen, 1981.

Institute of Linguists (undated): “Diploma in Translation: Marking Guidelines” (manuscript), Institute of Linguists, London.

Laurian, A-M. (1984): “Machine Translation: What type of Post-editing on what type of documents for what type of users”, in *Coling 1984,* Association for Computational Linguistics, Morristown, N.J.

Pierce, John R., and Carroll John B., et al. (1966): *Language and Machines – Computers in Translation and Linguistics* (= ALPAC report), National Academy of Sciences, National Research Council, Washington D.C.

Nagao, M., Tsujii, J., and Nakamura, J. (1985): “The Japanese Government Project for Machine Translation” in *Computational Linguistics,* Vol. 11, Nos 2-3, 1985.

Newmark, P. (1981): *Approaches to Translation,* Pergamon Press, Oxford, England.

Newmark, P. (1988): *A Textbook of Translation,* Prentice Hall, Hemel Hempstead, UK.

Nida, E.A. (1964): *Toward a Science of Translating,* E.J. Brill, Leiden, the Netherlands.

Sager, J.C. (1989): “Quality and standards – the evaluation of translations”, in Picken, *C.:The Translator's Handbook,* Aslib, London, pp. 91-102.

Snell-Hornby, M. (1988): *Translation Studies: an Integrated Approach,* J. Benjamins, Amsterdam and Philadelphia.

Translation Bureau of the Secretary of State of Canada (1990): “Guide-lines and Procedures for the Evaluation of Translations Done on Contract” (manuscript), Secretary of State of Canada, Translation Bureau, Canada.

Van Slype, G. (1982): “Conception d’une méthodologie générale d’évaluation de la traduction automatique", in *Multilingua* 1-4, pp. 221-237.