[*Mechanical Translation*, vol.3, no.3, December 1956; pp. 73-80]

Some Psychological Methods for Evaluating the Quality of Translations †

George A. Miller and J. G. Beebe-Center, Harvard University, Cambridge, Massachusetts

The excellence of a translation should be measured by the extent to which it preserves the exact meaning of the original. But so long as we have no accepted definition of meaning, much less of exact meaning, it is difficult to use such a measure. As a practical alternative, therefore, we must search for more modest, yet better defined, procedures. The present article attempts to survey some of the possible methods: One can ask the opinion of several competent judges. Or, given a translation of granted excellence, one can compare test translations with this criterion by a variety of statistical indices. Or a person who has read only the translation may be required to answer questions based on the original. The characteristic advantages and disadvantages of each method are illustrated by examples.

ONE HEARS it said that MT is currently rather crude, but that workers in the field are striving to improve and refine their translations. A brief encounter with the unedited output of an automatic dictionary is sufficient evidence of the tremendous range of quality between the simplest mechanical 'translation' and the product of a skilled, human translator. The question is whether this intuitive judgment of the quality of a translation can be made more precise by any psychological techniques of scale construction.

A scale of the quality of translations should be reliable, valid, objective and easy to use. In addition to these general desiderata for all scaling procedures, there are certain special features that this particular scale should have. For example, it should be applicable to any translation, whether produced by a machine or by a human translator. This feature would enable us to compare the output of a particular machine to the output of a human who had had a known number of years of study in the foreign language. Furthermore, the scale should be applicable to translations from or into any language whatsoever, and so should not take advantage of any characteristics peculiar to a given language, say English — Whether or not a single scale can apply to all languages and still make linguistic sense is a debatable question. And, preferably, the scale should be unidimensional, so that different translations could be compared with respect to a single 'figure of merit'. Finally, we would like to have one or more cutoff points indicated along the scale; "completely unusable," "useful for scanning as to subject matter", "useful after post-editing", "immediately readable, " and "suitable for publication" are some criteria that we might hope to locate along the scale.

All these features would be desirable, but it is not obvious at present that they can be achieved.

Subjective Scaling

Perhaps the most direct approach is to give both the original passage and the translation to be tested to a person who understands both languages and to ask him to assign a number between 0 and 100 to the translation, where 0 means that it is equivalent to no translation at all and 100 means the best imaginable translation. This method fails the criterion of objectivity, of course, and cannot be applied when a polyglot is not available to judge, but we expected to be able to map out the general territory in this way and to use subjective ratings

† Preparation of this paper was supported under Contract AF 33 (038) — 14343 between the U.S. Air Force and Harvard University and appears as Report Number AFCRC —TN—56 — 61, ASTIA Document Number AD 98823. Reproduction for any purpose of the U.S. Government is permitted.

We would like to acknowledge the assistance of Peter Aldin, Martha Taylor, Soon Duk Koh and Elizabeth Friedman.

as a criterion against which to test various other scaling techniques.

In a short exploratory study, however, we obtained somewhat confusing results. We found much disagreement among different raters. Perhaps we should have used foreign language teachers as our judges, for they probably have skill in grading that ordinary, bilingual persons do not seem to have, but we did not anticipate that the ratings would be so difficult.

For the purposes of this study, we selected four summaries of articles from the journal Acustica, two in German and two in French. The journal also gave an English translation, so we had the work of a theoretically competent translator to use for comparison. (The published translations were not the best possible, but they represent the sort of thing that is available in the current scientific literature.) Then we prepared mechanical translations, simulating by hand the possible operation of an automatic dictionary. Each word of the original text was written on a card. These cards were then alphabetized, and on the reverse side we listed the possible English equivalents in approximately the order of their frequency of occurrence, as well as we could judge it on intuitive grounds. From this pack we then constructed six different translations: (1) the first English alternative was chosen from each card; (2) an editor selected the best of the first two alternatives from each card, making his selection in complete ignorance of the other alternatives or the original passage; (3) an editor selected the best one from all the alternatives on each card, still in complete ignorance of the original passage; (4) an editor rewrote the English passage from a knowledge of only the first alternative on each card; (5) an editor rewrote the English passage from a knowledge of only the first two alternatives on each card; and (6) an editor rewrote the English passage from a knowledge of all the alternatives on each card, but without seeing the original passage. In all cases, these editors were monolingual Americans with no linguistic training. The first three procedures did not lead to grammatical English, of course, so we obtained a fairly wide range of quality by these procedures. These six translations, together with the translation taken from the journal and the original passage, were presented to judges who rated them on a scale from 0 to 100.

As a sample of the sort of materials produced, consider a single sentence taken from a French passage:

Original. Il résulte de ceci qu'une atmosphère stratifiée doit toujours réfléchir et donc produire des échos.

(1) He result of this which a atmosphere stratified must always to think and therefore to produce of the echoes.

(2) It results from this which a atmosphere stratified must always to reflect and therefore to produce of the echoes.

(3) It results from this that a atmosphere stratified must always reflect and therefore produce echoes.

(4) The result of this is that in a stratified atmosphere, one must always think of the echoes that are produced.

(5) It results from this that a stratified atmosphere must always reflect and therefore produce echoes.

(6) It results from this that a stratified atmosphere always reflects and therefore always produces echoes.

Published translation. It follows from this that a stratified atmosphere should reflect sound and produce echoes under all circumstances.

A similar sample taken from one of the German passages is the following:

Original. Bei beliebiger Impulsform ergibt sich das Faltungsprodukt aus Membran- und Impulsform.

(1) By any form of the impulse yields -self the products of the folding out membrane- and form of the impulse.

(2) By any form of the impulse yields the products of the folding out membrane- and form of an impulse.

(3) By any form of the impulse yields the products of the folding out membrane- and form of an impulse.

(4) Any form of the impulse is yielded by the interaction of the bending out of the membrane and the form of the impulse.

(5) The impulse in any form yields the products of the folding-out membrane and the form of an impulse.

(6) Any form of the impulse yields the products of the membrane-folding.

Published translation. With a given impulse form one obtains a resultant effect of the shapes of the impulse and of the disk.

Table I

Mean Ratings of Quality of Seven Translations

Method of Translation	French I	French. II	French Mean	German I	German II	German Mean
(1)	21.9	28.2	25.1	27.1	22.2	24.7
(2)	35.5	30.1	32.8	21.6	37.0	29.3
(3)	47.3	27.7	37.5	13.3	29.0	21.2
(4)	38.2	70.1	54.2	45.6	31.8	38.7
(5)	90.5	80.4	85.5	24.0	34.0	29.0
(6)	75.9	54.3	65.1	45.5	77.5	61.5
Published Translation	89.5	80.1	84.8	77.0	75.5	76.3

When the seven translations were given to subjects to judge, of course, no information was supplied as to the method of translation. It is interesting to note that supplying several alternative English equivalents seems to be more useful in translating from French than from German, but this judgment is based upon only these four samples of about 75 words each.

Eleven judges were used for the French passages and ten for the German. The judges were able to speak the language from which the translations came, but had no linguistic training; they were instructed to compare each translation with the original and to take time enough to be sure of their judgments. The means of their ratings are summarized in Table I.

There was so much disagreement among the judges (which was reflected in their bitter comments about the difficulty of their task) that even the means reveal only very general trends. These trends are clearer if we pool the data further, as in Table II.

From Table II we see that far more success is possible with French than with German, and that selective editing helps a little but not so much as complete rewriting. These conclusions are intuitively correct, and it would be disappointing indeed if they failed to appear. The error variance is so large, however, that these conclusions are barely significant.

We were slightly surprised that rewriting made as much difference as it did, since the people who rewrote had essentially the same information about the original passage as was contained in the selectively edited translations. The superiority of the rewritten translations indicated that the judges relied rather heavily upon the grammaticalness of the translation in reaching their decisions. In order to check this notion, we asked another group of subjects to act as judges, giving them the same instructions as before except that they were not shown the original French or German passages. Their ratings correlated closely with the original ratings, especially for the translations from German. It seems, therefore, that people will not regard favorably an ungrammatical translation even though they are able to understand it correctly.

Table II

Mean Ratings for Three MT Procedures for French and German

Method	French	German
No editing (1)	25.1	24.7
Selective editing (2-3)	35.2	25.3
Rewriting (4-6)	68.3	43.1
Means	53.4	38.6

We can conclude that a simple word-for-word substitution, method (1), is not satisfactory, but that an automatic dictionary combined with rewriting is a fairly satisfactory solution for translating from French into English. The problems with German are more difficult and seem to require that the machine recognize syntactic features. These conclusions, however, are of less immediate importance to us than the conclusions we can draw about this method of estimating the quality of translations: (a) The method is subjective; (b) Raters dislike the task; (c) There is considerable error variance, so that many judges are needed in order to obtain reliable means; (d) The literary skill of the rewriter is an important factor in the ratings; (e) An attempt should be made to obtain more experienced judges — either language teachers or professional translators.

Word Scores

Another way to approach the problem is to consider what a grader does when he evaluates a pupil's translation. Introspective reports indicate that he looks for two kinds of errors: (1) errors in vocabulary and (2) errors in construction. It is difficult to make these introspections more precise, for vocabulary and syntax are complexly intertwined. Nevertheless, it seems worthwhile to try.

The fact that a grader can recognize errors at all implies that he must have some personal standard against which he compares the student's work. In its most rigid form, this might consist of his own written translation; more often it is probably a rather vague set of translations that would be about equally acceptable. In order to imitate his procedures, therefore, we should have one or more explicit translations, written out in advance, that we will use as criteria. The task is then to obtain some objective measure of the relation between the test translation and the criteria.

Given a test and a criterion translation, the simplest thing to try first is to ask if they use the same words. That is to say, a score can be given by taking the number of words in the test translation which are duplicates of words in the criterion translation and then expressing this number as a fraction of the total number of words in the criterion translation. This method ignores the order in which the words are written. As an illustration:

Original: La maison se trouve à droite.

Criterion: The house is on the right.

Test: The house leans to the right.

From the criterion translation an alphabetical check list of words is prepared and the words in the test translation are checked against it:

house	1 √	
is	1	
on	1	Score = 4/6 = 0.67
right	1 √	
the	2 √√	

A number of exploratory experiments have been conducted with this method, using translations produced by students attempting to pass their language examinations in French or German and by competent translators. These studies have explored various possibilities, but none of them has been followed up with large amounts of data. Disregarding levels of significance, the studies can be summarized as follows:

(1) Five subjects with a good knowledge of both languages translated a sentence from German into English. These translations, all assumed subjectively to be 'good', were evaluated against a criterion translation. The scores ranged from 0. 73 to 0. 86. With students whose knowledge of German ranged from low to high, scores ranged from 0.19 to 0.70. For three persons with little knowledge of German, the mean score was 0.31. Four persons with a relatively good knowledge of German had a mean score of 0.65.

(2) One passage was translated from French into English by a simple word-for-word substitution, taking the first English equivalent that occurred in a French-English dictionary. The score for this translation was 0.40.

(3) One person who knew no Turkish but was familiar with the general subject matter translated a short, technical passage from Turkish into English. No dictionary was used. The score for a language as little related to English as this was 0.20. The fact that the score was not zero is due to the occurrence of common words in the two languages.

(4) In order to study the variability of the score, eleven French sentences were translated with a mean score of 0.65. The standard deviation was found to be 0.12.

(5) Seven translations of two German sentences were made by students. These were scored and the scores were compared with scores given by a grader on a longer passage containing these same sentences and also with scores on an 'objective test' of German language ability and achievement. The three measures of the students' ability were in close agreement.

(6) Since the use of a particular criterion translation may seem rather arbitrary, the check lists from six different criterion translations were combined and used to score the students' translations. With one criterion translation, there was a ceiling of about 0.86 and a mean of 0.50. When six criterion translations were combined, the ceiling rose to about 0.95 and the mean increased to 0.58. No significant changes in the rank order of the test translations resulted from this broader definition of the scoring criterion.

(7) When successive pairs of words, instead of individual words, were used to construct the check list, the scores were lower but were linearly related to the scores for individual words. With sequences of three successive words used to construct the check list, scores were very low and discrimination appeared to be lost.

(8) A word-for-word substitution of Korean equivalents for English words was made with ten sentences totalling 171 words in length. The Korean words, in the English order, were given to three Korean students at Harvard. They were asked to rewrite the sentences in Korean, ignoring as best they could their knowledge of English. Their rewritten sentences were then scored against a criterion prepared by an experienced translator. The three scores averaged 0.49. However, if differences in inflection are ignored and the word is considered correct if the root is identical, the average was 0.75. It is very likely, however, that the subjects' familiarity with English was a considerable aid to them.

(9) These same sentences were then translated again, this time using some simple rules for pre-editing the English. (a) Articles were omitted; (b) Idioms were underlined; (c) When 'of' occurred in a possessive phrase, the order of the words was inverted; and (d) When 'to' occurred in an infinitive construction, it was indicated. With this pre-editing, the word-for-word translation was repeated. The two sets of sentences, translated with and without pre-editing, were given to two groups of 31 students each in the Kyung-Bock High School, Seoul, Korea, and they were asked to rewrite them into intelligible Korean sentences. Their sentences were then scored against the criterion translation. The average score without pre-editing was 0.125; with pre-editing, 0.218. These scores are probably too low; the students were being given instruction during the summer vacation because of their poor school records.

These studies support some general comments. For human translators, a simple measure of correspondence of vocabulary correlates rather well with a subjective evaluation of the quality of the translation; a student who has achieved a given level of competence in vocabulary has probably achieved a corresponding level of competence in grammar, so the vocabulary measure will be correlated with any other measure of quality. For MT, however, the correspondence is not so close. It is possible to imagine a mechanical translation that is completely unintelligible yet contains most of the correct words. That is to say, the vocabulary measure is necessary but not sufficient. Nevertheless, we have been pleasantly surprised that so mechanical and simple a procedure gives us any discrimination at all.

Word-Order Scores

In order to supplement the simple vocabulary score, we would like to have some indicant of the syntactical adequacy of the translation. Before bringing to bear the more sophisticated concepts of modern linguistics, we decided to try the simplest possible comparison with a criterion translation. The simplest method we could think of was to compare the order of the words which were common to the test and the criterion translations. For example:

Criterion: The young boy walked fast.

Test: The fast boy had walked.

From the criterion translation a check list is again prepared, but this time the ordinal position of each word is indicated:

	Position in Criterion	Position in Test
boy	3	3 √
fast	5	2
the	1	1 √
walked	4	5 √
young	2	

The word score is 4/5 = 0.80, when scored as before. If we consider the four shared words, we find that the three checked words correspond as to order. Thus the word-order score can be stated as 3/4 = 0.75.

Thirteen people, whose knowledge of French varied from low to high, were given four 300-word French passages to translate. These translations were scored by the word-order method and also by a more subjective technique, with a grader scoring errors in words and in phrases. Furthermore, each person took two forms of an objective examination in French language achievement.

The word-order scores ranged from 0.20 to 0.72. The error scores given by the grader ranged from 1.6 to 24.4. The objective examination scores ranged from 252 to 750 (where 250 is chance performance). Thus all three measures discriminated among the translators. The average correlation between word-order scores and error scores was about 0.70, and between the word-order scores and the objective examination scores was about 0.60.

The reliability of the word-order score is reasonably good and could probably be improved by lengthening the passages. The correlation with error scores and objective examinations provides evidence for some degree of validity, at least for human translators. This technique is useful to discriminate against very poor translations, but the present evidence indicates that it may not discriminate accurately in the range that might be labelled 'good' to 'excellent'.

A slightly more sophisticated and less mechanical way to get at the syntactic aspects has been used by Koh in the Korean studies. A scoring key is constructed in advance by noting which words modify other words in the original English passage. If the rewritten Korean translation contains this same relation, one point is given. When the rewritten translations produced by the Korean high school students were scored by such a key, they obtained an average score of 8.5% on the passages without pre-editing and 23. 3% with pre-editing. The method is rather arbitrary, inasmuch as the experimenter must select in advance those syntactic relations for which credit will be given, and it is less mechanical than the word-order score, since it requires some intelligent judgment both in constructing the key and in doing the scoring. Nevertheless, it is a technique that deserves further exploration.

These methods involving a statistical comparison of the test translation with a criterion translation are certainly effective at the lower end of the scale. Whether the statistical net can be woven fine enough to catch the subtle shades of meaning that differentiate between 'acceptable' and 'good' or 'excellent', however, is still an open question.

Measures of Transmitted Information

One goal, although an unrealistic one, that we might hope to attain in translation is reversibility. That is to say, we could recover the original passage exactly by translating back again. We do not usually aspire to this goal, because it is not necessary to recover exactly the original passage. Various alternative wordings may be adequate for purposes of communication; so we hope merely to land somewhere inside this set of acceptable alternatives. When we translate we hope that something will remain invariant under translation. This something might be called the meaning or it might be called the information. Since techniques for estimating amounts of information have been developed, this line of thought leads to the suggestion that we should attempt to compare different translations to see how much information they have in common.

The method we have explored is one developed by Claude Shannon for estimating the redundancy of printed texts. Subjects guess repeatedly at successive letters, advancing to letter n + 1 after they have correctly guessed letter n. Shannon has shown how to estimate the amount of information, in bits per letter, from the frequency distribution of correct responses on the first, second, third, etc., guess. In fact, Miller and Friedman[2] have found that it is not necessary to obtain repeated guesses, since the amount of information per letter can be estimated rather closely from the percentage of times the first guess is correct. The relation is H = 5Q, where H is the number of bits per letter, and Q is the probability of being wrong on the first guess.

---

1. Shannon, C.E., "Prediction and Entropy of Printed English", Bell Syst. Tech. J. 1951, 30, 50-64.

2. Miller, G.A., and Friedman, E.A., "The Reconstruction of Mutilated English Texts", Information and Control, 1957 (in press).

The strategy we have used involves an approximation to the information formula,

$$T = H(x) - H_y(x),$$

where T is the amount of information common to x and y; H (x) is the amount of information in x; and $H_y(x)$ is the amount of information in x when y is known. Now suppose that x and y are two alternative translations of the same passage. We can estimate H(x) by asking a subject to guess successive letters according to Shannon's technique. Then we can take another subject and show him translation y; with y available to him, he now proceeds to guess successive letters in x, and so gives us an estimate of $H_y(x)$. Assuming the two subjects to have identical guessing habits, the difference between these two measures should give us an estimate of the amount of information common to the two translations. If one translation is a criterion translation, the value of T should be high when the test translation contains essentially the same information, and low when it contains relatively little of the same information as the criterion.

In a preliminary study we found that T averaged 0.8 bits per letter for two 'good' translations of a given sentence and 0.05 bits per letter for one 'good' and one 'poor* translation. Although these results indicate that the method may be feasible, it is laborious and time-consuming; we have not explored a wide variety of conditions in this way and will probably not do so unless it becomes of some further theoretical interest. It does have the slight advantage that the measure is given in bits per letter, which may be more meaningful to computer designers than some more arbitrary scale.

Reading Comprehension Tests

A possible criticism of the methods discussed so far is that they are too much concerned with the small details of a translation and too little concerned with the general purpose of making translations in the first place. The purpose, of course, is communication. The translation should be judged successful if this purpose is achieved.

In ordinary situations outside the psychologist's laboratory, we have a simple check on whether we have communicated successfully. We ask questions. For example, after a series of communicative acts that he calls 'lectures', a teacher will evaluate his success by a procedure that he calls an 'examination'. If the recipients of a message can answer correctly questions which they could not answer before they received the message, we conclude that the communication was successful.

One way to apply this technique is in the form of commands that must be carried out by some gross, bodily behavior. A more convenient way is to ask questions that can be answered verbally. For example, in order to evaluate the readability of a particular passage, psychologists give the reader a few minutes to study it and then ask him a series of questions ranging from very simple to very difficult. Once a set of passages has been standardized for readability on a large sample of readers, it can be used to measure the reading skill of other individuals. Such a set of passages with related questions is called a 'reading comprehension test'. It should be relatively straightforward to apply this same technique to measure the comprehensibility of a translation.

The translation to be tested would be presented to a person along with a list of questions that he must answer about the meaning of the passage. These questions should be simple enough that an intelligent person equipped with a good translation could answer them all, yet difficult enough that a person with no translation could not answer any of them. We have hesitated to adopt this approach because the phrasing of the questions requires much skill and the test should be standardized on relatively large groups of subjects.

For example, the subject might be presented with the following word-for-word translation of a German passage:

> The theory the passage of sound through plates is — for even waves and bounded bundle — in such form given that the relation with it the free waves of the plate in appearance steps. Cremer's conception the total number of passages as 'coincidences' the falling in wave with it free waves of the plate, certain exceptions hereof and the influence a final cross section of the wave are discusses. The conclusions are experimental with it ultra-sound on aluminum plate proven.

Then he would be confronted by questions like the following:

1. What does the form of the theory reveal?
2. What was done with the conclusions?
3. What kind of incident sound was studied analytically?

4. What kind of incident sound was studied experimentally?
5. Was Cremer's theory accepted without qualification?
6. What did Cremer think was coinciding?

Although these questions have not been tested in any way, it is hoped that they will be difficult to answer until you have read the following alternative translation:

> The theory of transmission of sound — plane waves and laterally bounded beams — through plates is given in a form which reveals the connection with the free waves in plates. Cremer's interpretation of total transmission as 'coincidence' of the incident wave with a free wave in the plate, certain exceptions from that representation, and the influence of the finite cross section of the beam are discussed. The conclusions have been examined experimentally on aluminum plates by ultrasonic waves.

This example should make clear the difficulties involved in formulating good questions. On the one hand, they should not be so specific as to require a particular word in answer, for this reduces to a vocabulary test. On the other hand they should not be so general that it is difficult to decide whether the answer is right or wrong. No doubt special passages would have to be constructed for the purpose; we have not yet undertaken this formidable task.

Syntactic Analyses

All of the scaling procedures discussed above are linguistically naive. We have been much impressed by the elegance of certain theories of grammar. For example, Z. Harris' constituent analysis should certainly yield some kind of measure of agreement between the true analysis and the constituents of the translation to be tested. However, these ideas have been difficult to apply because the translations produced by some of the simpler mechanical procedures are so bad that it is impossible to say what the constituents are. Such analysis is easier if the translation is grammatical.

Ideas concerning the degree of grammaticalness of a passage are suggested in the work of A. N. Chomsky. For example, if words are classified into syntactic categories, we might ask how often ungrammatical sequences of categories occur. As a variable we could examine the degree of precision of the syntactic classification. A very grammatical translation would have only permissible sequences even with the most refined analysis of categories, whereas an ungrammatical translation might not have only permissible sequences until the categories were reduced to something as crude as Noun, Verb, Adjective, and X, where X represents everything else. This is a forbidding task to undertake, however, and does not get at the question of whether the translation, grammatical or not, carries the same meaning as the original. Indeed, much syntactic analysis carefully avoids any contamination with semantics.

We have assumed, therefore, that such analyses are much more important for workers trying to develop translating machines than for those who would like to evaluate the finished product.

Our studies have not explored the closely related problem of measuring the "translatability" of the original passages. We have observed, of course, that with respect to English, French is more translatable than German. But there are many other differences. The literature in any given language is not uniformly translatable, and some schemes for MT may succeed with one author and fail with another. For example, a passage which is well written in the original language will usually be more translatable than a poorly written passage. Or, again, a passage written by a person who knows no English will usually be harder to translate into English than something written in the same language by a person whose first language was English. Only a large sample of different materials in the source language can inform us on this question, and it is impractical to generate such a sample by manual simulation. Thus there are important aspects of the evaluation problem that cannot be studied satisfactorily until the machines are running.