[Mechanical Translation, vol. 8, No. 2, February 1965]

Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments

by Sheila M. Pfafflin*, Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

This paper discusses the results of an experiment designed to test the quality of translations, in which human subjects were presented with IBM-produced machine translations of several passages taken from the Russian electrical engineering journal Elektrosviaz, *and with human translations of some other passages taken from* Telecommunications, *the English translation of* Elektrosviaz. *The subjects were tested for comprehension of the passages read, and were also asked to judge the clarity of individual sentences. Although the human translations generally gave better results than the machine translations, the differences were frequently not significant. Most subjects regarded the machine translations as comprehensible and clear enough to indicate whether a more polished human translation was desirable. The reading comprehension test and the judgment of clarity test were found to give more consistent results than an earlier procedure for evaluating translations, since the questions asked in the current series of tests were more precise and limited in scope than those in the earlier scries.*

In view of the considerable effort currently going into mechanical translation, it would be desirable to have some way of evaluating the results of various translation methods. An individual who wishes to form his own opinion of such translations can, of course, read a sample, but this procedure is unsatisfactory for many purposes. To indicate only one difficulty, individuals vary widely in their reactions to the same sample of translation. However, a previous attempt by Miller and Beebecenter[1] to develop a more satisfactory approach gave discouraging results. When ratings of the quality of passages were used, it was found that subjects had considerable difficulty in performing the task, and were highly variable in their ratings; while information measures, which were also used, proved very time-consuming. Furthermore, neither of these methods provided a direct test of the subject's understanding of the translated material.

The present study explored two other approaches to the evaluation problem, namely, reading comprehension tests, and judgments of the clarity of meaning of individual sentences. The approach through testing of reading comprehension provides a direct test of at least one aspect of the quality of translation. Judgments of sentence clarity do not, but they are likely to be simpler to prepare and may have applicability to a wider range of material. Both types of tests might therefore be useful for different evaluation problems if they proved to be effective. While the previous results with a rating technique are not encouraging for a judgment method, the assignment of one rating of over-all quality to a passage is a fairly complex task. We hoped that by asking subjects to judge sentences rather than passages, and to judge for clarity of meaning only, rather than quality generally, the subjects' task would be simplified and the results made more reliable.

Test Materials and General Procedures

In these evaluations, passages translated from Russian into English by machine were compared with human translations of the same material. Technical material was chosen for the subject matter, since the major efforts in machine translation have been directed towards it; the specific field of electrical engineering was selected because a large number of technically trained subjects were available in it.

Eight passages were selected from a Russian journal of electrical engineering, *Elektrosviaz.* These passages were used in the reading comprehension test and also provided the sentences for the clarity rating tests. Insofar as possible, bias toward particular subject matter was avoided by random selection of the volume and page at which the search for each passage started. However, in order to make up a satisfactory comprehension test, it was desirable to avoid material involving graphs or equations. The result is that the majority of the passages come from introductions to articles. The translated passages vary in length from 275 to 593 words.

The machine translations of these passages were provided by IBM and were based on the Bidirectional Single-Pass translation system developed there by G. Tarnawsky and his associates. This system employs an

* The author wishes to express her appreciation to Mrs. A. Werner, who prepared translations for preliminary tests and advised on preparation of the final text, and to D. B. Robinson, Jr., L. Rosier, and B. J. Kinsberg for their contributions to selection of passages and preparation of questions used in the reading comprehension tests.

analysis of the immediate linguistic environment to eliminate the most common ambiguities in the Russian language and to smooth out the English translation. The only alterations in the computer output were the substitution of English equivalents for a few Russian words not translated, and minor editing for misprints. The human translations used were taken from the journal *Telecommunications,* the English translation of *Elektrosviaz.*

Members of the Technical Staff at Bell Telephone Laboratories with a background in electrical engineering were used as subjects in all of the experiments to be described. They were randomly selected from the available subjects.

Reading Comprehension Tests

PREPARATION OF THE READING COMPREHENSION TEST

The questions for the test were made up from the original Russian passage by two electrical engineers. They used multiple-choice questions with four possible answers. The number of questions per passage varied from four to seven, for a total of 41 questions. The same questions were used with human and machine translations of a given passage.

Prior to their use in the comprehension test, 27 subjects answered the questions without reading any translation in order to determine how well they could be answered from past knowledge alone. The average number of correct answers was 14.6, somewhat higher than the 10.25 correct answers to be expected from guessing alone. The figure obtained from the guessing test should therefore be taken as the basis for comparison, rather than the theoretical chance level.

FIRST READING COMPREHENSION TEST

Method

Sixty-four subjects were used in the experiment. Each subject answered questions on four human and four machine translations of different passages. An 8 by 8 randomized Latin Square was used to determine the order in which the passages were presented to the subjects. Four sequences of human and machine translations were imposed on each row of the Latin Square; HHHHMMMM, MMMMHHHH, HMHMHMHM, MHMHMHMH. Two subjects received each combination of passage and HM order. Practice effects were thus controlled for both types of translations and passages, and the effect of changing to the other type of translation after different amounts of practice could be observed.

Procedure

Subjects were run in groups of up to four. They were allowed to spend as much time reading each passage as they chose, but were not allowed to refer back to the passage once they had begun to answer questions about it. Opinions of the translations were obtained from some subjects following the test.

Results

The average number of questions answered correctly is given in Table 1. Performance following either type of translation is clearly above the guessing level. The difference in number of correct responses for human and machine translations is significant at the 0.01 level, as determined by the sign test.*

TABLE 1

Mean Number of Questions Correctly Answered, Both Reading Comprehension Tests

	Human	*Machine*
RCT 1	32.7	28.4
RCT 2	34.1	32.2

The individual passages differed somewhat in difficulty, but there was no apparent effect of the position of the passage in the test, as such, on number of correct responses. Neither was there any over-all difference between the four patterns of ordering human and machine translations. However, the number of errors decreases slightly for those machine translated passages which are preceded by other machine translated passages (see Table 2). This decrease is just significant at the 0.05 level, according to the Friedman analysis of variance.* No practice effect is apparent for passages translated by humans.

TABLE 2

Mean Number of Errors by Order of Occurrence of Translation Methods, Reading Comprehension Test 1

	Position			
Method	*1*	*2*	*3*	*4*
Human	70	63	59	74
Machine	112	95	107	87

The amount of time which the subjects spend reading the two types of passages is given in Table 3. The subjects spent more time in reading the machine trans-

TABLE 3

Mean Reading Time, in Minutes per Passage, by Order of Occurrence of Translation Method, Reading Comprehension Test 1.

	Position				
Method	*1*	*2*	*3*	*4*	*Mean*
Human	3.7	3.7	3.8	3.5	3.7
Machine	5.1	5.2	4.3	3.8	4.6

* vide reference 2.

lations than they did the human translations. This measure shows a practice effect in the case of the machine translations, though not for human translations. The difference in reading time between the human and machine translations is significant at the .001 level, according to the sign test, and the decreasing amount of reading time taken by the subjects is significant at the .05 level according to the Friedman nonparametric analysis of variance.*

In addition to the measures of time and number of questions correctly answered, 43 of the subjects gave their opinion as to whether the machine translations were: (1) adequate in themselves, (2) adequate as a guide for deciding whether to request a better translation, or (3) totally useless. Sixteen subjects also gave their opinion of the human translations; the results are shown in Table 4.

TABLE 4

Proportion of Opinions on Adequacy of Translations in the Three Categories, Reading Comprehension Test 1

		Opinion	
Method	*Adequate*	*Guide*	*Useless*
Human	.87	.13	.00
Machine	.10	.86	.04

The comments made by subjects judging the human translations as only partially adequate suggest their judgments were made less favorable by the fact that the passages were not complete articles. Presumably this factor also affects the judgments of machine translation, though there is no direct evidence from comments. The comments most often made about the machine translations suggest that they required more attention and rereading than the human translations. Some comments also indicated that subjects were disturbed by failure to select prepositions and articles appropriately.

SECOND READING COMPREHENSION TEST

Method

The materials, design and other procedures used in this test were similar to those of the first reading comprehension test, with the following changes. Timing data and opinion were not recorded. Thirty-two subjects were used, and the sequences of human and machine passages alternated for all subjects. Subjects were not only allowed as much time as they liked to read the passages, but were allowed to refer back to them in answering the questions.

Results

The number of correct responses for human and machine passages is shown in Table 1. Performance is better for both machine and human passages than it was in the first test, and the difference between the two is no longer significant.

DISCUSSION OF THE READING COMPREHENSION TESTS

In considering the results of the reading comprehension tests, perhaps the most striking feature is the relatively small difference in the number of correct responses for the two types of translations. Although the difference between them in this regard is significant when the subjects are required to answer from memory, it is not large, and it becomes insignificant when subjects are allowed to refer back to the passages in answering the questions. This result stands in contrast to the opinions collected about these translations, which showed that most subjects considered the human translations adequate, but considered the machine translations adequate only as a guide in deciding whether a better translation was needed. This result may reflect, in part, the emotional reactions of subjects to the grammatical inadequacies of the machine translations. It probably also reflects differences in the effort required to understand the two types of translations.

Thus, while these results indicate that a good deal of information is available in machine translations, they are also consistent with the view that it is less readily available than in human translations. They also suggest that practice with the machine translations can improve readers' ability to understand them, which is consistent with the subjective opinions of those who have used machine translations.

Judgment of Clarity Tests

In the following series of tests, subjects were requested to state whether they considered individual sentences translated by the two methods to be clear in meaning, unclear in meaning, or meaningless. The unclear category was further defined to include sentences which could be interpreted in more than one way, as well as sentences for which a single interpretation could be found, but with a feeling of uncertainty as to whether it was the intended interpretation.

Subjects in the first study judged the sentences in paragraphs. It was intended as a preliminary to judgments of sentences separated from their context in paragraphs, and therefore a relatively small number of subjects were run. However, the data have been included since they provide some information about the effect of context on the judgments.

The other two tests differ from the first study in that each sentence appeared on a separate card, in random order, so that context effects were largely absent. In one of these tests, the same subjects judged sentences translated by both methods; in the other the same subjects judged only one type of translation.

* vide reference 2.

CONTINUOUS TEXT TEST

Materials

The eight passages used in the reading comprehension tests were divided into two sets of four, and the sentences in each passage were numbered. The same sets were used for both human and machine translations. Subjects received either all human or all machine translations.

Procedure

Sixteen subjects divided into two groups of eight judged the machine translated passages. Each group judged one of the sets of four passages.

Eight additional subjects divided into two groups of four judged the sentences in the equivalent passages translated by humans. The subjects indicated their answers on separate answer sheets. They were run in groups up to four.

SEPARATE SENTENCES TEST, MIXED TYPES

Materials

Sixty sentences were randomly selected from the passages used in the reading comprehension test. The human and machine translations of these sentences were typed on IBM cards. Underneath the sentences were the numbers 1, 2, or 3, which the subjects circled to indicate the category in which they placed the sentence. Each subject was also given a separate card which stated the meanings of the three categories.

Design

The sentences were divided into two groups of thirty each. The human translations from one group of thirty sentences were then combined with the machine translations of the other thirty sentences to form two sets of sixty sentences. Twenty five subjects judged each set; different subjects were used for the two sets.

Procedure

The subjects were run in groups of up to eight. They were first read instructions which explained the judgments they were to make; these instructions emphasized to the subjects that they were to judge on meaning, not grammar. They then proceeded through the decks of sentences at a self-paced rate. The sentences were in a different random order for each subject.

SEPARATE SENTENCE TEST, SEPARATE TYPES

Materials

The same sixty sentences used in the previous separate sentence test were used here.

Design

The sixty sentences, all in machine translation, were judged by twenty five subjects. Twenty five different subjects judged the sentences in human translation.

Procedure

The procedure was the same as in the mixed-types test.

RESULTS OF THE JUDGMENTS OF CLARITY TESTS

The results of all three tests are shown in Table 5. The ratings of the sixty sentences used in the separated sentence tests are shown separately for the context test. The results suggest that there is no effect due to the presence or absence of context on judgments of sentences translated by humans, but that judging them along with machine translations increases the proportion of clear judgments assigned to them. In the case of the machine-translated sentences, there appears to be both a context effect, and a depressing effect upon the judgments when they are made along with judgments of human translated sentences. When the sign test was applied to the differences in number of clear and unclear judgments of individual sentences under the two separate sentence conditions, they were found to be significant (.01) level). Similar tests of the differences between machine translated sentences when judged in context and out of context in the absence of sentences translated by humans were significant at the .05 level.

TABLE 5

Proportions of judgments in different categories for judgment experiments (C = clear, UC = unclear, NM = no meaning. In eases where two groups of Ss judged under the same conditions, proportions are averages of both. Separate sentences, context, arc judgments in context for these sentences which were used in separate sentence tests).

	Human			*Machine*		
Test	C	UC	NM	C	UC	NM
Context:						
All Sentences	.80	.16	.04	.65	.27	.08
(Separate Sentences)	.79	.16	.05	.68	.25	.07
Separate Sentences:						
Same Ss	.91	.08	.01	.39	.40	.21
Different Ss	.77	.20	.03	.49	.33	.18

The distribution of the responses is also markedly different for the two types of translations. Figure 1 gives the distribution of the sentences according to the number of subjects who assigned the sentence to a given category. The distribution of responses varies more for the sentences translated by machine than for the sentences translated by humans.

In order to get a single number which characterized each sentence, the numerical values 1, 2, and 3 were

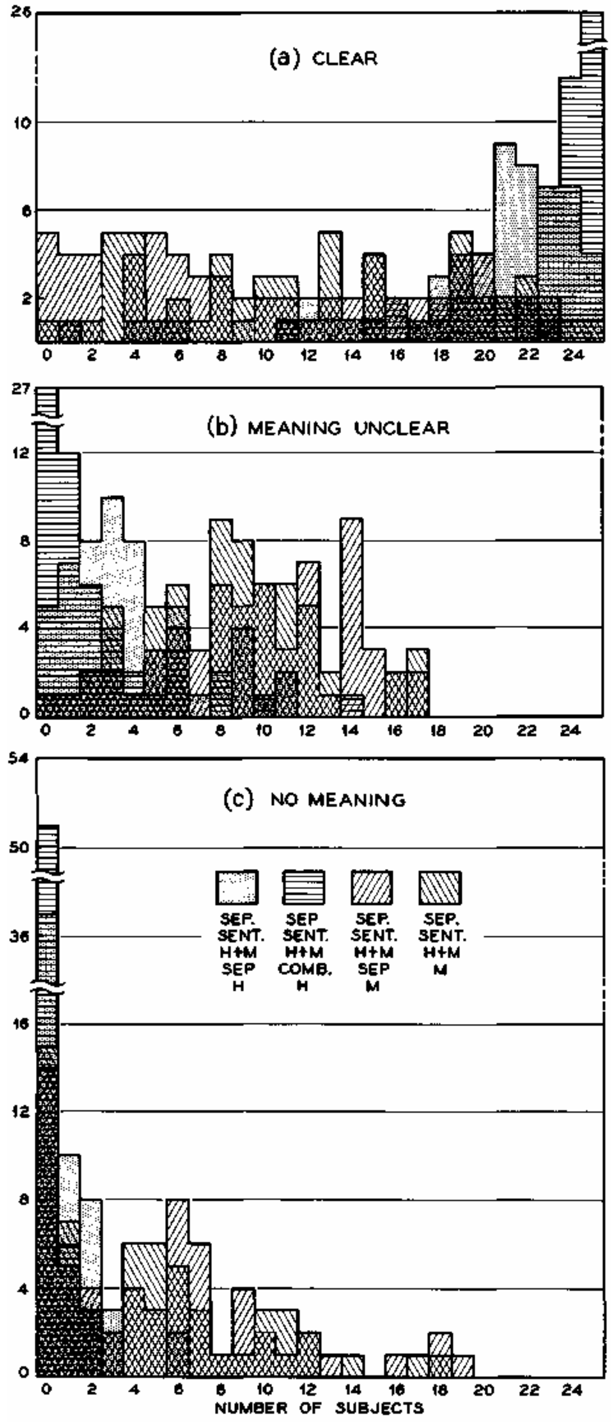

Figure I

Distribution of the sentences according to the number of subjects assigning a sentence to a given category. The abscissa shows the number of subjects who made a given type of response to a given sentence. The ordinate shows the number of sentences which received this pattern of response from the subjects. The three categories of response are shown separately. Method of translation and judgment condition are indicated by different patterns.

assigned to the categories and the values of the judgments assigned to each sentence were summed. The frequency with which different subjects used the categories is clearly different, so that if one assumes that the subjects have an underlying ordering for these sentences, while differing in the point at which they shift from one type of response to the next, the summing of the responses given to each sentence should give a reasonable indication of the rank order of that sentence relative to others which are judged. The resulting scale values provide good discrimination between the machine translated sentences. They also appear to be reliable; the Spearman rank order correlation between the scale values assigned to machine translated sentences judged in combination with human translations and those judged separately is over .9 for both groups of subjects. The judgments do not, however, discriminate among the sentences translated by humans, except in the case of a few sentences which were judged low in meaning.

Efforts were made to relate the scale values of the sentences to some other measures which might be thought to indicate quality of the translation. No relation was found to the length of the sentence, when the difficulty of the sentence in the original translation was taken into account by ratings of the human translations. Nor was a relation found between number of words which were identical or similar in the two types of translations. There appeared to be a low correlation between the number of errors which subjects made in the reading comprehension tests and the average scale values of the sentences in these passages, but it did not reach a satisfactory level of significance.

DISCUSSION OF THE JUDGMENTS OF CLARITY TESTS

The finding that mixing the types of translations during judging affects both types of translations, while loss of context in a paragraph affects only machine translations, is hardly surprising. The range of values of a set of stimuli along a judged continuum is known to affect the distribution of responses for all stimuli in the set. The additional effect of context, on the other hand, would be expected to appear only if many of the sentences were unclear when judged out of context, which is the case only for the machine translations. The context effect for such sentences supports the earlier evidence from the reading comprehension tests that information is less readily available in these machine translations.

The general lack of success in relating the judgments to some other possible indices of quality is also not surprising, since these indices, with the exception of the reading comprehension scores, were very simple measures, and previous work* had already indicated that such measures were unlikely to be useful. They

* vide reference 1.

were tested here to insure that the judgments were not simply covering the same ground as these obvious measures, at greater cost. It would, of course, have been helpful if it had been possible to demonstrate a clear relation between judgment scores and reading comprehension scores. However, a number of factors militated against the likelihood of doing so in these experiments. First, fewer than half of the sentences in the reading comprehension tests were rated by enough subjects to provide scale values. Furthermore, performance on the reading comprehension tests is also a function of passage difficulty and question difficulty, and considerably more data would be needed adequately to separate out these effects from that of method of translation.

One other aspect of the data should be commented on, and that is the relative reliability of the rating method used here, compared with the high variability which the previous investigators reported with rating methods. The difference is probably due in part to the question asked. Subjects were asked to judge sentences on one dimension only, clarity, and were not asked to give over-all estimates of quality, which would take into account such questions as style and grammar, and which could therefore lead to highly variable judgments.

The reliability of this method may also be due in part to the fact that the sentences were rated in isolation, without context; the judgments which were obtained from the sentences in context appear to show more intersubject variability than sentences rated in isolation, though it has not been possible to measure this difference quantitatively in a satisfactory manner. However, since it is reasonable to assume that context interacts with both sentences and subjects, it would not be surprising if judgments in context were more variable than judgments out of context. While for some purposes, tests without context may be undesirable, it would seem that for purposes of deciding whether differences exist between two methods of translation, out of context judgments may be entirely adequate, and perhaps even superior to judgments in context, for, questions of reliability aside, the structure of the material translated may convey sufficient information to mask real differences between the methods.

General Discussion

The amount of effort involved in preparation and administration is one important consideration for an evaluation method. The sentence judgment method is easier than the reading comprehension test, if the effort involved in developing the test is considered, and it appears to provide a reasonably reliable estimate of relative sentence clarity. The absolute value of these judgments is, of course, subject to the types of biases already noted. It would, however, appear to be a fairly simple method for determining whether or not two methods of machine translation differ from each other in the number of understandable sentences which they produce.

On the other hand, this judgment method does not provide a direct measurement of the usefulness of a translation. Possibly, despite the problems raised by response biases, some relations to direct performance measures could be worked out, at least sufficiently to give a crude measure of predictability from sentence judgments. However, in the absence of some demonstrated relationships, it would appear undesirable to depend on sentence judgments alone.

Another consideration is that of sensitivity. It is fairly clear that the sentence judgment method has at least the potential for more sensitivity than this particular reading comprehension test, since the judgment results show a much larger range than the reading comprehension test results. Two points should be noted here. First, it may be possible to develop more sensitive comprehension tests. Second, the judgment method may be too sensitive for some uses. That is to say, it may show statistically significant differences between translation methods which do not differ in any important way in acceptability to the user.

Even tests of reading comprehension, however, directly test only one aspect of a translation's adequacy. Since it can be expected that machine translations would frequently be read for general information, rather than to obtain answers to specific questions, the question arises as to what extent the results of this test can be generalized to other uses of machine translations. Much controversy exists over the adequacy of multiple choice questions to test general understanding, as distinct from recall of specific facts, and this paper will not attempt to add anything to the already considerable amount of discussion on this topic. However, as far as the evaluation of translations goes, providing readers with sufficient information to enable them to answer multiple-choice questions about its contents would appear to be a minimum requirement for a useful translation, and hence can provide a baseline, even while it is recognized that such a test may not be sensitive to more subtle factors which would be important in some uses.

Ideally, of course, one would wish to have one or more tests that would evaluate all aspects of translation quality, but at the present time this goal is visionary; it is not even possible to state with any certainty just what all these aspects are. The problem may be partially solved by changes in the translations themselves. If the point is ever reached where subjects who read both human and machine translations of the same material are unable to distinguish between them, and bilingual experts cannot decide which type gives a more accurate translation, the problem of evaluation will simply disappear. And if, as has been suggested, translation methods can be developed which give grammatical, though not necessarily accurate, translations, the nature of the evaluation problem will be radically changed. At the present time, however, a combination of several methods, including the two investigated here, would appear likely to be of some use.

Summary and Conclusions

Evaluation of the quality of machine translations by means of a test of reading comprehension and by judgments of sentence clarity, was investigated. Human translations and IBM machine translations of passages from a Russian technical journal were used as test materials. Performance on the reading comprehension test was better when human translations were used, but the difference was not large, and was significant only when the subjects were not allowed to refer back to the passages when answering the questions. The subjects generally felt that the machine translations were adequate as a guide to determine whether a human translation was desired, but inadequate as the sole translation. When the subjects judged sentences selected from the passages for clarity of meaning, machine translated versions were in general considered less clear than human translated versions. The judgments were found to discriminate among the machine translated sentences, though not among the sentences when translated by humans. While tests of reading comprehension provide a more direct measure of the usefulness of translations than do judgments of sentence clarity, the latter approach is simpler, and may be more sensitive. Both methods therefore may be of value in evaluating machine translations.

References

1. Miller, G. A., and Beebecenter, J. G., "Some Psychological Methods for Evaluating the Quality of Translations," *Mechanical Translation,* 1958, 3, 73-80.
2. Siegel, S., *Non-Parametric Statistics,* New York, McGraw-Hill, 1956.