Appendix 10

An Experiment in Evaluating the Quality of Translations

This experiment* was designed to lay the foundations for a standard procedure for measuring the quality of scientific translations, whether human or mechanical. There have been other experiments on this problem [e.g., G. A. Miller and J. G. Beebe-Center, Mechan. Transl. , 3, 73 (1958); S. M. Pfafflin, Mechan. Transl. *8,* 2 (1965)], but their methods for evaluating translations have been too laborious, too subject to arbitrariness in standards, or too lacking in reliability and/or validity to become generally accepted. The measurement procedure developed here gives promise of being amenable to refinement to the point where it will meet the requirements of relative simplicity and feasibility, fixed standards of evaluation, and high validity and reliability.

A detailed report of this experiment will be submitted for publication elsewhere; the present brief report will serve to indicate the general nature of the measurement procedure and some of the chief results.

THE MEASUREMENT PROCEDURE

It was reasoned that the two major characteristics of a translation are (a) its intelligibility, and (b) its fidelity to the sense of the original text. Conceptually, these characteristics are independent ; that is, 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; this would be likely to occur, however, only in cases where the original had low intelligibility.

Essentially, the method for evaluating translations employed in this experiment involved obtaining subjective ratings for these two characteristics—intelligibility and fidelity—of sentences selected

* Conducted by John B. Carroll with funds provided by the Automatic Language Processing Advisory Committee.

randomly from a translation and interspersed in random order among other sentences from the same translation and also among sentences selected at random from other translations of varying quality. When a translation sentence was being rated for intelligibility, it was rated without reference to the original. "Fidelity" was measured indirectly: the rater was asked to gather whatever meaning he could from the translation sentence and then evaluate the original sentence for its "informativeness" in relation to what he had understood from the translation sentence. Thus, a rating of the original sentence as "highly informative" relative to the translation sentence would imply that the latter was lacking in fidelity.

All ratings were made by persons who were specially selected and trained for this purpose. There were two sets of raters. The first set of raters (called here "monolinguals" for convenience) consisted of 18 native speakers of English who had no knowledge of the language of the original (Russian, in this case). They were all Harvard undergraduates with high tested verbal intelligence and with good backgrounds in science. In rating "informativeness" these raters were provided with carefully prepared English translations of the original sentences, so that in effect they were comparing two sentences in English—one the sentence from the translation being evaluated, and the other the carefully prepared translation of the original.

The second set of raters ("bilinguals") consisted of 18 native speakers of English who had a high degree of competence in the comprehension of scientific Russian. Their ratings of the intelligibility of the translation sentences may well have been influenced by their knowledge of the vocabulary and syntax of Russian; at any rate, no attempt was made to prevent them from using such knowledge. To rate "informativeness," they made a direct comparison between the translation sentence (in English) and the original version.

All ratings were made on nine-point scales that had been established by the writer prior to the experiment by an adaptation of a psychometric technique known as the method of equal-appearing intervals. Thus, points on these scales could be assumed to be equally spaced in terms of subjectively observed differences. In the case of the intelligibility scale, each of the nine points on the scale had a verbal description (see Table 4). The same was true of the "informativeness" scale except that verbal descriptions were omitted for a few of the points (sec Table 5). In this way each degree on the scales could be characterized in a meaningful way. For example, point 9 on the intelligibility scale was described

TABLE 4. Scale of Intelligibility

9—Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.
8—Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or midly unusual word usage that could, nevertheless, be easily "corrected."
7—Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.
6—The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Postediting could leave this in nearly acceptable form.
5—The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.
4—Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated.
3—Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.
2—Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.
1—Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

as follows: "Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities." Point 5 (the midpoint of the scale): "The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly 'noise' through which the main idea is still perceptible."

PREPARATION OF TEST MATERIALS AND COLLECTION OF DATA

The measurement procedure was tested by applying it to six varied English translations--three human and three mechanical —

TABLE 5. Scale of Informativeness

(This pertains to how informative the original version is perceived to be after the translation has been seen and studied. If the translation already conveys a great deal of information, it may be that the original can be said to be low in informativeness relative to the translation being evaluated. But if the translation conveys only a certain amount of information, it may be that the original conveys a great deal more, in which case the original is high in informativeness relative to the translation being evaluated.)

9—Extremely informative. Makes "all the difference in the world" in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation.)
8—Very informative. Contributes a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the reader's impression of the meaning intended, although not so much as to change or reverse the meaning completely.
7—(Between 6 and 8.)
6—Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader "on the right track" as to the meaning intended.
5—(Between 4 and 6.)
4—In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.
3—By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different "twist" to the meaning conveyed by the translation. It adds no new information about sentence structure, however.
2—No really new meaning is added by the original, either at the word level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended.
1—Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced.
0—The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.

of a Russian work entitled Mashina i Mysl' (Machine and Thought), by Z. Rovenskii, A. Uemov, and E. Uemova (Moscow, 1960). These translations were of five passages varying considerably in type of content. (All the passages selected for this experiment, with the original Russian versions, have now been published by the Office of Technical Services, U.S. Department of Commerce, Technical

Translation TT 65-60307.) The materials associated with one of these passages were used for pilot studies and rater practice sessions; the experiment proper used the remaining four passages.

In preparing materials for the rating task, 36 sentences were selected at random from each of the four passages under study. Since six different translations were being evaluated, six different sets of materials were prepared (in two forms, one for the monolinguals and one for the bilinguals) in such a way that each set contained a different translation of a given sentence. In this way no rater evaluated more than one translation of a given sentence. Each set of materials was given to three monolinguals and to three bilinguals; thus, there were 18 monolinguals and 18 bilinguals. Each rater had 144 sentences to evaluate first for intelligibility and then for the informativeness of the original (or the standard translation of it) after the translation had been seen. The raters required three 90-min sessions to complete this task, dealing with 48 sentences in each session. The raters were not informed as to the source of the translations they were rating, although they were told that some had been made by machine.

Before undertaking this task, the raters attended a 1-hr session in which they were given instruction in the rating procedures and required to work through a 30-sentence practice set.

During the rendering of ratings for intelligibility, the raters held stopwatches on themselves to record the number of seconds it took them to read and rate each sentence.

RESULTS

The results of the experiment can be considered under two headings: (a) the average scores of the various translations, and (b) the variation in the scores as a function of differences in sentences, passages, and raters.

Table 6 gives the over-all mean ratings and time scores for the six translations, arranged in order of general excellence according to our data.

Consider first the mean ratings for intelligibility by the monolinguals. Translation 1, a published human translation that had presumably been carefully done, received the highest mean rating, 8.30, on the scale established in Table 4. But 8.30 is still appreciably different from the maximum possible mean rating of 9.00, and it is evident that not even this "careful" human translation was as good as one might have expected. Furthermore, the mean rating of Translation 1 is not significantly different from that of Translation 4 (8.21), a "quick" human translation made by rapid dictation

procedures. The mean ratings of Translations 1 and 4 do, however, differ significantly from the mean rating (7.36) of Translation 2, another "quick" human translation. It may be concluded that the measurement procedure studied here is sensitive enough to differentiate among human translations.

A similar remark may be made about the sensitivity of this procedure to differences in the intelligibility of machine translations. Translations 7 and 5 were shown to be significantly more intelligible, on the average, than Translation 9.

Of most current interest, however, are the results having to do with the comparison of the human and the machine translations. Machine translations 7, 5, and 9 received mean ratings, respectively, of 5.72, 5.50, and 4.73. A scale value of 5 refers to a translation in which "the general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands ..." All these machine translations are significantly less intelligible, on the average, than any of the three human translations. As machine translations improve, it should be possible to scale them by the present rating procedure to determine how nearly they approach human translations in intelligibility.

The monolinguals' mean ratings on "informativeness" (reflecting the lack of fidelity of the translations) show an almost perfect inverse relationship to the mean ratings on intelligibility, and they differentiate the various translations in the same way and to the same extent. This result means that in practice, when ratings are averaged over sentences, passages, and raters, "intelligibility" and "fidelity" are very highly correlated. The detailed results of this study show that only in the case of a few particular sentences do the mean ratings of intelligibility and informativeness convey different information.

Furthermore, the mean reading times per sentence show almost precisely the same pattern of results as the ratings. In fact, the mean reading times are linearly related to the mean ratings, a result that supports the conclusion that the points on the rating scales are evenly spaced.

The results from the ratings by bilinguals contribute nothing more to the differentiation of the translations than is obtainable with the monolinguals' ratings. Bilinguals' intelligibility ratings of the translations are slightly (and significantly) higher, on the average, than those of the monolinguals, and correspondingly, their informativeness ratings are slightly lower. Yet, they took significantly longer to read and rate the sentences. Apparently their knowledge of Russian caused them to work harder on trying to understand the translations. One is inclined to give more credence to the results

from the monolinguals because monolinguals are more representative of potential users of translations and are not influenced by knowledge of the source language. It is also to be noted that the data from the monolinguals differentiate the translations to a somewhat greater extent than do the data from the bilinguals.

The results concerning the differences in ratings due to differences in sentences, passages, and raters can now be considered. (The detailed tables of these results are omitted here to save space.) The more important results may be summarized as follows:

1. The results do not differ significantly from passage to passage; that is, on the average the various passages from a given translation receive highly similar ratings. For intelligibility ratings, however, there is a small but significant interaction between translation and passage, indicating that translations are to some extent differentially effective for different types of content. (This interaction effect is present both for human and for machine translations.)

2. There is a marked variation among the sentences. In fact, as may be seen from Figure 1, there is some overlap between sentences from human translations and from mechanical translations; or, in other words, there are some sentences translated by machine that have higher ratings than some other sentences translated by human translators, even though, on the average, the human-translated sentences are better than the machine-translated ones. These results imply that in order to obtain reliable mean ratings for translations, a fairly large sample of sentences must be rated.

3. Variation among raters is relatively small, but it is large enough to suggest that ratings should always be obtained from several raters—say at least three or four.

CONCLUSION

This experiment has established the fact that highly reliable assessments can be made of the quality of human and machine translations. In the case of the six particular translations investigated in the study, all the human translations were clearly superior to the machine translations; further, some human translations were significantly superior to other human translations, and some machine translations were significantly superior to other machine translations. On the whole, the machine translations were found to fall about at the midpoint of a scale ranging from the best possible to the poorest possible translation.

What is still needed, however, is a system whereby any translation can be easily and reliably assessed. The present experiment has determined the necessary parameters of such a system.

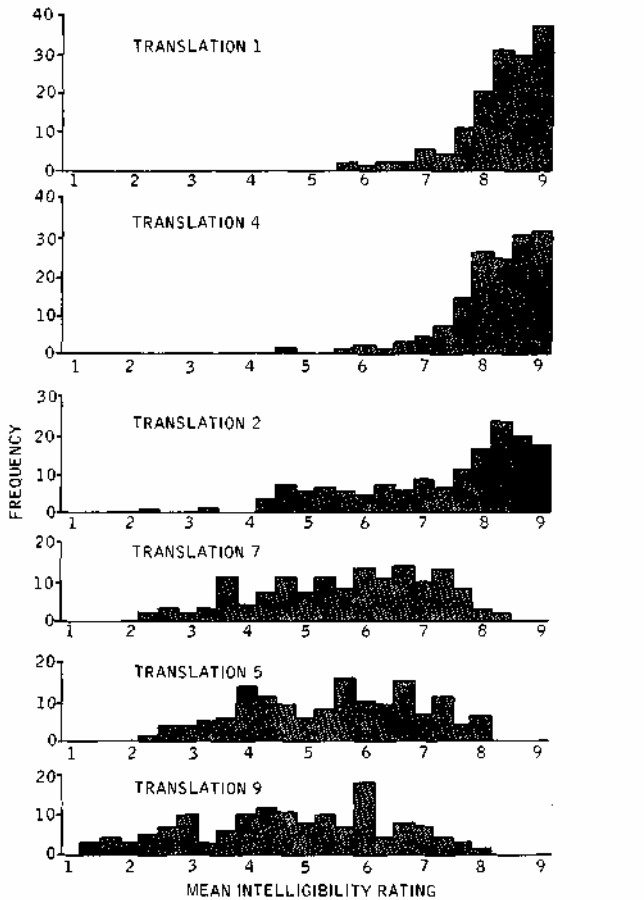

FIGURE 1. Frequency distribution of monolinguals' mean intelligibility ratings of the 144 sentences in each of six translations. Translations 1, 4, and 2 are human translations; Translations 7, 5, and 9 are machine translations.

TABLE 6. Evaluation of Translations: Over-all Mean Ratings and Time Scores from "Monolingual" (M) and "Bilingual" (B) Raters[a] (3 raters x 36 sentences x 4 passages = 432 observations underlying each mean)

Translation Number	Description	Mean Ratings Intelligibility M	Intelligibility B	Informativeness M	Informativeness B	Mean Reading Times per Sentence (sec) M	Mean Reading Times per Sentence (sec) B
1	"Careful," published human translation	8.30	8.37	1.95	1.72	9.13	10.09
4	"Quick" human translation	8.21	8.25	1.85	1.47	9.21	11.54
2	"Quick" human translation	7.36	7.67	3.03	2.43	12.59	13.53
7	Machine translation, Program B 2nd Pass	5.72	5.86	4.28	4.19	18.89	20.50
5	Machine translation, Program A	5.50	5.59	4.41	3.88	18.98	20.42
9	Machine translation, Program C 1st Pass	4.73	5.14	5.34	5.09	23.96	23.75

[a]The translations are listed in order of decreasing general excellence according to the results presented here. The brackets indicate results of the application of the Newman-Keuls multiple-range test of the significance of the differences of the rank-ordered means in each column. Any two means embraced within a given bracket are not significantly different at the 0.01 level; any two means not embraced within one bracket are significantly different at the 0.01 level. There are several cases in which the above listing entails reversals of the order of means, but in no case are the means involved significantly different from each other.