[Mechanical Translation and Computational Linguistics, vol.9, nos.3 and 4, September and December 1966]

An Experiment in Evaluating the Quality of Translations

by John B. Carroll,* Graduate School of Education, Harvard University

To lay the foundations for a systematic procedure that could be applied to any scientific translation, this experiment evaluates the error variances attributable to various sources inherent in a design in which discrete, randomly ordered sentences from translations are rated for intelligibility and for fidelity to the original. The procedure is applied to three human and three mechanical translations into English of four passages from a Russian work on cybernetics, yielding mean scores for the translations. Human and mechanical translations are clearly different in over-all quality, although substantial overlap is noted when individual sentences are considered. The procedure also clearly differentiates within sets of human translations and within sets of mechanical translations. Results from the two scales are highly correlated, and these in turn are highly correlated with reading times. A procedure in which highly intelligent "monolingual" raters (i.e., without knowledge of the foreign language) compare a test translation with a carefully prepared translation is found to be more reliable than one in which "bilingual" raters compare the English translation with the Russian original.

Introduction

It would be desirable, in studies of the merits of machine translation attempts, to have available a relatively simple yet accurate and valid technique for scaling the quality of translations. It has also become apparent that such a technique would be useful in assessing human translations. The present experiment seeks to lay the foundations for the development of a technique.

There have been several other experiments in measuring the quality of mechanical translations,[1,2] but the procedures proposed in these experiments have generally been too laborious, too subject to arbitrariness in standards, or too lacking in validity and/or reliability to constitute a satisfactory basis for a standard evaluation technique. For example, Pfafflin's method requires that a reading-comprehension test be constructed for each translation that is to be evaluated, and thus it allows latitude for considerable variance in the difficulty of the test questions and permits sliding standards in the scale of measurement.

The present experiment develops a method that appears to meet requirements of high validity, high reliability, fixed standards of evaluation, and relative simplicity and feasibility.

The method is based on the following considerations:

1. The evaluation of the adequacy of a translation must rest ultimately upon subjective judgments, that is, judgments resulting from human cognitions and intuitions. (If any objective measurements directly applicable to the translations themselves were available—say, some form of word-counting—they could presumably be used in the production of translations; hence, use of such objective procedures in the evaluation of translations could lead to circularity.)

2. If sufficient care is taken, procedures utilizing subjective judgments can be devised that attain acceptable levels of reliability and validity and that yield satisfactory properties of the scale or scales on which measurements are reported.

3. Certain types of objective measurement of the behavior of human beings in dealing with translations can be useful in providing evidence to corroborate the validity of subjective measurements, but they cannot serve as the sole basis for an evaluation procedure because they do not directly indicate adequacy of translation.

In order to obtain subjective measurements of known reliability and validity, it was believed necessary to do the following:

1. Obtain measurements of all the dimensions thought logically necessary and essential to represent the adequacy of a translation—namely, intelligibility and fidelity—as will be explained below.

2. Develop rating scales with *(a)* relatively fine graduations (nine points rather than three or five as used in some previous studies); *(b)* equality of units estab-

* I wish to thank Mr. Richard See of the National Science Foundation, Dr. A. Hood Roberts of the Automatic Language Processing Advisory Committee, National Academy of Sciences-National Research Council, and Dr. Ruth Davis of the Department of Defense, for help in obtaining and selecting the Russian translations that were to be evaluated; Dr. J. Van Campen and Dr. Charles Townsend of the Department of Slavic Languages and Literatures, Harvard University, for help in constructing superior translations of the Russian; Dr. Maurice Tatsuoka of the University of Illinois, and Dr. J. Keith Smith of the University of Michigan, for advice on statistical analyses; Dr. Mary Long Burke Betts for assistance in data collection and statistical computations; and Miss Marjorie Morse, Jr., for clerical assistance. The facilities of the Harvard Computing Center were used. Author's address after February 1, 1967: Senior Research Psychologist, Educational Testing Service, Princeton, New Jersey 08540.

lished by a standard psychophysical technique, and if possible validated with reference to a correlated variable; and *(c)* verbal descriptions of the points on the scale so that measurements could be directly interpreted.

3. Divide the translations to be measured into small enough parts (translation units) so that a substantial number of relatively independent judgments could be obtained on any given translation, and so that the variance of measurement due to this kind of sampling could be ascertained.

4. Provide a collection of translation units that would be sufficiently heterogeneous in quality to minimize the degree to which the judgments on the evaluative scales would be affected by varying subjective standards (a rectangular distribution of stimuli along the scales being regarded as the ideal).

5. Take account of, and where possible investigate, variables in the selection of judges that might affect the reliability, validity, and scaling of measurements.

6. Train judges carefully for rating tasks demanded of them.

7. For each translation unit, obtain judgments from more than one rater so that the variance of measurement attributable to raters could be ascertained.

Background

The present experiment was made possible through the efforts of representatives of the Joint Automatic Language Processing Group, who made the arrangements whereby a total of nine varied translations of the same work—*Mashina i Mysl' (Machine and Thought),* by Z. Rovenskii, A. Uemov, and E. Uemova (Moscow, 1960)—became available. Four of these translations were human, five were by machine; of these translations, only six were complete, however, and for the purposes of the present study comparisons were made only for passages selected from these. With the assistance of Dr. Ruth Davis, Department of Defense, Mr. Richard See, Office of Science Information Services, National Science Foundation, and also of Dr. A. Hood Roberts, executive secretary of the Automatic Language Processing Advisory Committee, the writer selected five passages of varied content, each containing at least fifty or sixty Russian sentences. One passage, drawn from the General Introduction to the book, was used for various pilot studies, rater training, etc., and will not be reported on. The other four passages, numbered 2, 3, 4, and 5, concerned the following subjects: (2) the technical prerequisites of cybernetics; (3) logic; (4) the origin of cybernetics; (5) characteristics of human behavior which cannot be reproduced by a machine. (All the passages selected for this experiment, with the original Russian versions, have now been published.[3]) The six translations that were involved in this experiment (aside from one other special translation that will be mentioned below) were coded as follows:

Translation No. 1: an allegedly "careful," published human translation
Translation No. 2: a rapid human translation, presumably done "at sight" by dictation
Translation No. 4: another rapid human translation, done by a different translator
Translation No. 5: a machine translation (Machine Program A)
Translation No. 7: a machine translation (Machine Program B, 2d Pass)
Translation No. 9: a machine translation (Machine Program C, 1st Pass)

Preparation of Material

The first step toward preparing the data for the experiment was to have each sentence of the Russian original typed on a 5 × 8-inch card; suitable identifying code numbers were placed on the back of each card. The corresponding material in each of the six translations was then identified and similarly typed on cards, one card for each translation. Russian sentences were identified in terms of the occurrence of full stops (periods) or question marks. In most cases, there was a one-for-one correspondence between sentences of the original Russian and of the translations, but occasionally the human translators made two or more English sentences out of a single Russian sentence, or, conversely, merged the content of two Russian sentences into one English sentence. In any case, the Russian sentence as defined by punctuation was the unit of analysis. There were occasional cases in which a translation for a given Russian sentence was either missing completely or given only in part through obvious carelessness, and in such cases all translations for the given sentence were eliminated from further consideration because the object of the study was to study the adequacy of translation *when a translation was available* (the carelessness of translators being regarded as something controllable by suitable administrative procedures). Sentences in which the Russian contained mathematical formulas or tabular material were also eliminated from consideration.

The rationale for choosing the sentence for the unit of analysis (implying that sentences would be considered out of context and in random order) was that it was thought that a minimum requirement on a translation would be that each sentence of a translation should convey at least the "core" meaning conveyed by the corresponding original when taken in isolation. Many translation sentences, of course, will convey more than this; that is, the translator will often use the total context of the passage in order to supply certain critical and needed meanings, for example, the gender of a pronoun left unspecified in the original. Likewise, it is sometimes legitimate for a translation to omit certain elements of meaning present in the original when the structure of the translation language does not demand that such elements be specified and when

they will be understood from the context. It was felt, however, that such minor discrepancies would balance out and would be taken account of by the raters in such a way as to introduce little if any error into the procedures that were developed.

For a reason that will become apparent later in connection with the total design of the study, it was found necessary to have translations of the Russian originals of whose quality one could be assured. Originally it had been thought that Translation No. 1 would serve this purpose, but careful inspection of this translation and comparison with the Russian original disclosed that it contained not only numerous minor blemishes in English phraseology but also a number of questionable and possibly misleading translations. Consequently, the services of Drs. Joseph Van Campen and Charles Townsend, both members of the Department of Slavic Languages and Literatures of Harvard University (and the latter a thoroughly experienced professional translator of scientific Russian), were obtained to make translations (using the complete context) of all five passages involved in the experiment. These translations were coded as Translation No. 0 and typed, sentence by sentence, on cards in the manner described previously.

Development of Rating Scales

The next step was to develop rating scales to measure any and all dimensions thought logically necessary and essential to represent the adequacy of a translation (apart from such mechanical considerations as legibility, completeness of graphics, etc.). Drawing on discussions of this matter in the meetings of the Automatic Language Processing Advisory Committee, the writer concluded that there were two such dimensions: intelligibility and fidelity or accuracy.

The requirement that a translation be intelligible means that as far as possible the translation should read like normal, well-edited prose and be readily understandable in the same way that such a sentence would be understandable if originally composed in the translation language. (In the case of translations of highly technical, abstruse, or recondite materials, this requirement means only that the material be intelligible to a person sufficiently acquainted with the subject matter or the level of discourse to be expected to understand it.)

The requirement that a translation be of high fidelity or accuracy has already been discussed, in part, in connection with justifying the sentence as the unit of analysis. In particular, it means further that the translation should as little as possible twist, distort, or controvert the meaning intended by the original. For the purposes of this experiment, the question of the fidelity of a translation was converted into the complementary question of whether the original could be found to contain no information that would supplement or controvert information already conveyed by the translation. It was assumed that *unjustified* supplying of information by a translation, as well as the omission or distortion of information, would contribute to lack of fidelity.

It was recognized that perfect fidelity of translation is not always possible, but it was assumed that raters of translations would take this fact into account in making their judgments.

In effect, then, fidelity of a translation was to be judged in terms of the "informativeness" of the original *relative to the translation.* In this way, the translation is being evaluated—not the original—since the judgments of the informativeness of the original are to be made only after the translation has been examined.

It should be noted that intelligibility (of the translation) and informativeness (of the original relative to the translation) are conceptually separable variables. For example, a translation could be perfectly intelligible, but the corresponding original could be completely "informative" in that it would completely contradict the translation; in this case, the translation would be maximally lacking in fidelity. The opposite case would be represented by a translation that was maximally unintelligible, matched by an original that was minimally informative; in this case, the original could be characterized as "bad, untranslatable text." Normally, however, it might be expected that intelligibility and informativeness would be in inverse relationship; that is, the original would be informative to the degree that the translation is lacking in intelligibility. (This proved to be the case in the great majority of instances, as will be shown below.)

The rating scale for intelligibility (see Table 1) was constructed in the following manner: Approximately two hundred sentences, consisting of nearly all the translations of the sentences in Passage 1, were sorted and re-sorted by the writer into nine piles of increasing intelligibility, so that the piles were as homogeneous as possible and the psychological distances between adjacent piles in the series appeared to be equal. (This is the standard psychophysical technique known as the method of "equal-appearing intervals.") There was no attempt to "force" the distribution of the cards, but, presumably because of the nature of the materials, the distribution was somewhat biased in the direction of an overrepresentation of higher intelligibility values as compared with the perfectly flat or rectangular distribution that might have been desired. Next, each pile was examined, and a verbal description was composed to characterize the degree of intelligibility that it represented. These verbal characterizations were discussed in one of the writer's advanced seminars in language measurement at Harvard University, and some modifications were made in the light of the resulting suggestions.

It may appear that the scale descriptions which resulted from this procedure incorporate some degree of

TABLE 1
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 mildly 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.

multidimensionality: In the upper end of the scale, differentiation between adjacent values depends largely on matters of style and word choice, whereas in the lower portion of the scale, it depends, rather, on matters of syntactical arrangement. The principal defense that can be made for treating several dimensions in a single scale is that the translations actually appear to arrange themselves along such a scale and the raters are able to make reliable global judgments on it.

The rating scale for informativeness (see Table 2) was constructed in a similar manner. The approximately two hundred sentences used in the previous sorting were paired up with their counterparts in the original (or, rather, in Translation No. 0, used as equivalent to the original because of the writer's relative lack of expertness in the Russian language) and sorted by the writer into nine piles of ascending degrees of "informativeness" of the original sentence relative to the translation sentence. Again, the method of equal-appearing intervals was used. It was found necessary to add a further pile at the lower end of the scale, with a scale value of zero, for the cases in which translations seemed justifiably to have supplied information, presumably from the total context, not present explicitly in the originals.

TABLE 2
SCALE OF INFORMATIVENESS*

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.

* 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.*

Selection of Raters

In order to study the effect of a critical variable in the selection of raters—their knowledge of the source language—the experiment was conducted in two parts. Part I employed eighteen male students in the junior (third) year at Harvard University, selected for their high verbal intelligence (Scholastic Aptitude Test [SAT] verbal scores 700 or greater) and for their interest and knowledge in science (since this was the general subject matter of the Russian work, the translations of which were to be evaluated). All were honors

majors in chemistry, biology, physics, astronomy, or mathematics. These students were screened to insure that they had no knowledge of Russian; in the rating task, they evaluated the informativeness of Translation No. 0 (as described above) relative to the translations under study. Part II utilized eighteen males selected for their expertness in reading Russian (generally, scientific Russian); most of these males were graduate students in Russian or teachers of Russian, and several were professional translators of scientific Russian. These persons were not screened for their knowledge or lack of knowledge of science, however.

All raters were native speakers of English. The screening of the raters in Part I of the experiment by means of SAT verbal scores was done to insure, as far as possible, that they would be suitably sensitive to the niceties of English phraseology and diction as well as to the intellectual content of the material. There was no such guaranty in the case of the raters used in Part II of the experiment, since it did not seem feasible to administer an intelligence test to them comparable to the College Entrance Examination Board Scholastic Aptitude Test. The fact that they were all university graduates experienced in problems of language translation, however, probably implies that their verbal intelligence scores would have averaged at a high level—perhaps as high as the average of the Part I raters. (For convenience in subsequent discussions, the raters in Part I are called "monolinguals," and the raters in Part II, "bilinguals" or "Russian readers.")

Organization of Materials to be Rated

In the main rating task, thirty-six sentences were selected at random from each of the four passages under study (Passages 2, 3, 4, 5). Since six different translations were being evaluated, six different sets of materials were made up for each part of the experiment (one series for monolinguals, one series for Russian readers) in such a way that each set contained a different translation of a given sentence, the sentence-translation combinations being rotated through the sets and presented in random order. This was done because it was considered imperative not to have a given rater rate a given sentence in more than one translation, since otherwise the ratings would lose independence. Furthermore, since the sentences were to be considered in isolation, they were presented in random order so as to reduce to practically zero any possibility that a rater could take context into account. Each of the six sets of material in each part of the experiment thus contained a total of 144 sentences, each sentence being represented by a particular translation and either the Translation No. 0 version (for the monolinguals) or the original Russian (for the bilinguals). In each part of the experiment, three raters were assigned to each of the six sets of material, so that there were eighteen raters in all in each part.

Further details concerning the organization of the materials are given in the following section.

Rating Procedures

Each set of material was divided into three subsets (I, II, III) of forty-eight sentences each, so that each rater could deal with his 144 sentences on three separate occasions called "main rating sessions," at least a day apart. Raters paced themselves and took, on the average, about ninety minutes per session. The order in which the subsets were dealt with by the raters was systematically permuted through the arrangements I, II, III; II, III, I; III, I, II. (If more than three raters had been used, more permutations could have been used.)

A day or so before any rater started on his three main rating sessions, he had a one-hour practice session in which he was introduced to the scales and the procedures (as described below) and given practice in applying them to thirty sentences (in various translations) selected from Passage 1. It is probable that the use of a rater-training procedure such as this is of importance in securing reliable and valid ratings, but it would be useful to check this point in further research.

The procedure for each of the main rating sessions was as follows: First, the rater evaluated the forty-eight translation sentences in the subset, one by one, for intelligibility according to the nine-point scale of Table 1. As he did so, he held a stopwatch and recorded both the intelligibility rating and the time (in seconds) that it took to read and rate each sentence. The time measurements were taken in order to obtain an objective correlate of the intelligibility ratings; both the time measurements and the intelligibility ratings are undoubtedly also correlated positively with the lengths of the translation sentences, but no account has been taken of these correlations in the present report because the length of a translation sentence relative to the original version was regarded as one of the variables involved in translation adequacy, and hence it was allowed to affect intelligibility ratings in an uncontrolled manner. (The validity of this assumption can be checked in further analyses of the data collected here.)

In this part of the procedure, that is, the rendering of intelligibility ratings and the associated time measurements, the rater saw *only* the translation sentences which were presented one sentence to a page in a loose-leaf format. (The pages were Xeroxed from the cards that had been prepared.)

Next, the rater turned to a portion of the loose-leaf book in which each successive page contained (by Xerox reproduction process) both a translation sentence and, just below it, a target sentence to be evaluated for informativeness according to the scale shown in Table 2. For monolinguals, of course, the target sentence was

in Translation No. 0, as described previously, while, for the bilinguals, the target was the original Russian sentence.

The materials were organized within each subset so that the order in which the sentence pairs were presented in this second part of the procedure was the same as that in which the translation sentences had been presented for the intelligibility ratings.

The procedures thus yielded three dependent variables: the intelligibility rating, an informativeness rating, and a time measurement for the intelligibility rating.

Externally, the rating for intelligibility was the same for the monolinguals and the bilinguals, in the sense that they were both rating precisely the same materials on the same scale and taking the same time measurements for their ratings. But since the bilinguals were familiar with Russian, it seemed unrealistic to expect them to evaluate the translations under the pretense that they did not know Russian, especially since the translations occasionally contained untranslated words (in transliteration) and other traces of the original, such as typical Russian word orders and idioms. Therefore, the Russian readers were told to evaluate the translation sentences from the standpoint of the maximal degree of intelligibility perceived in them, utilizing whatever ingenuity in comprehension they had as a result of their knowledge of Russian.

Results

The main results of the experiment are shown here, first, as a series of six analysis-of-variance tables (one for each of three dependent variables in each part of

TABLE 3

MEAN SQUARES AND TOTAL SUMS OF SQUARES IN VARIANCE ANALYSES FOR THREE DEPENDENT VARIABLES, PARTS I ("MONOLINGUAL") AND II ("BILINGUAL") OF THE EXPERIMENT ON EVALUATION OF TRANSLATIONS

Source	Number of Degrees of Freedom	Intelligibility		Informativeness		Reading Time Scores	
		I	II	I	II	I	II
Translations (T)	5	994.93**	899.04**	881.48**	929.30**	15,937**	13,678**
Passages (P)	3	5.32$^{n.s.}$	12.11$^{n.s.}$	3.87$^{n.s.}$	19.62$^{n.s.}$	646$^{n.s.}$	1,259$^{n.s.}$
Sentences (S) within P	140	10.67**	9.21**	21.57**	12.87**	690**	733**
T × P	15	12.23**	7.34**	11.00$^{n.s.}$	11.89*	182$^{n.s.}$	291*
T × S within P	700	3.79**	3.26**	8.00**	6.25**	173**	165**
Within cells (raters)	1,728	1.41	1.75	3.08	3.27	142	93
Total sum of squares	2,591	11,760	11,238	18,520	16,711	547,248	456,224

Note.—Symbols indicate significance levels of the F-ratios corresponding to the given mean squares with appropriate error terms as specified in the text: $**p < .01$; $*p < .05$; $^{n.s.}\ p > .05$ (not significant).

TABLE 4

EVALUATION OF TRANSLATIONS: OVER-ALL MEAN RATINGS AND TIME SCORES FOR PARTS I ("MONOLINGUAL") AND II ("BILINGUAL") OF THE EXPERIMENT*
(3 Raters × 36 Sentences × 4 Passages = 432 Observations underlying Each Mean)

Translation Number and Description	Mean Ratings: Intelligibility I	Mean Ratings: Intelligibility II	Mean Ratings: Informativeness I	Mean Ratings: Informativeness II	Mean Reading Times per Sentence (sec.) I	Mean Reading Times per Sentence (sec.) II
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, 2d Passage	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 Passage	4.73	5.14	5.34	5.09	23.96	23.75

* 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 .01 level; any two means *not* embraced within one bracket are significantly different at the .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.

TABLE 5

EXPECTED VALUES OF MEAN SQUARES FOR A THREE-FACTOR PARTLY HIERARCHAL EXPERIMENT

Source of Variation	*df*	Expected Value of Mean Square
A (translations)	$p-1$	$\sigma_e^2 + n\sigma_{ac}^2 + nr\sigma_{ab}^2 + nqr\sigma_a^2$
B (passages)	$q-1$	$\sigma_e^2 + np\sigma_c^2 + npr\sigma_b^2$
C within B (sentences within passages)	$q(r-1)$	$\sigma_e^2 + np\sigma_c^2$
AB (translations × passages)	$(p-1)(q-1)$	$\sigma_e^2 + n\sigma_{ac}^2 + nr\sigma_{ab}^2$
A × C (C within B) (translations × sentences within passages)	$q(p-1)(r-1)$	$\sigma_e^2 + n\sigma_{ac}^2$
E (within raters for a given translation sentence)	$pqr(n-1)$	σ_e^2

p = No. of translations (a fixed factor).
q = No. of passages (a random factor).
r = No. of sentences (a random factor).
n = No. of raters for a given translation sentence (a random factor).

Source: Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill Book Co., 1962, p. 189.

the experiment) contained in Table 3, and second, as a series of mean over-all ratings and time scores for the six translations, shown in Table 4. (Since passages did not differ significantly, separate data for passages are not given.)

The analysis-of-variance tables of Table 3 reflect the design of the study, in which (in each part of the experiment) groups of sentences in different translations rated by different sets of raters are "nested" within passages (Winer, 1962, p. 189, Table 5, 12-4).[4] The statistical model for the experiment is shown as Table 5. Since only the translation effect is fixed, the error term for translations is translations × passages; for passages, it is sentences within passages; for translations × passages, it is translations × sentences within passages. The within-cells mean square is the error term for sentences within passages and for translations × sentences within passages. It has been assumed, for convenience, that the rater effect is a completely random one. (Data are available to show that the rater effect is comparatively small.)

For all dependent variables, the translation effect is highly significant, a fact that indicates that the rating technique used here reliably differentiated at least some of the various translations. The passages do not, however, differ significantly over the whole set of data, although for some of the dependent variables there is a significant interaction between translation and passage. This may be interpreted to mean that the translations are differentially effective for the passages. This is particularly true for the intelligibility variable, where the interaction is highly significant for both parts of the experiment. The time scores and informativeness variables showed a barely significant ($p < .05$) translations × passages interaction for the Russian readers, but not for the monolinguals.

Sentences within passages is in every case a highly significant effect, as is also the interaction between translations and sentences within passages. These results

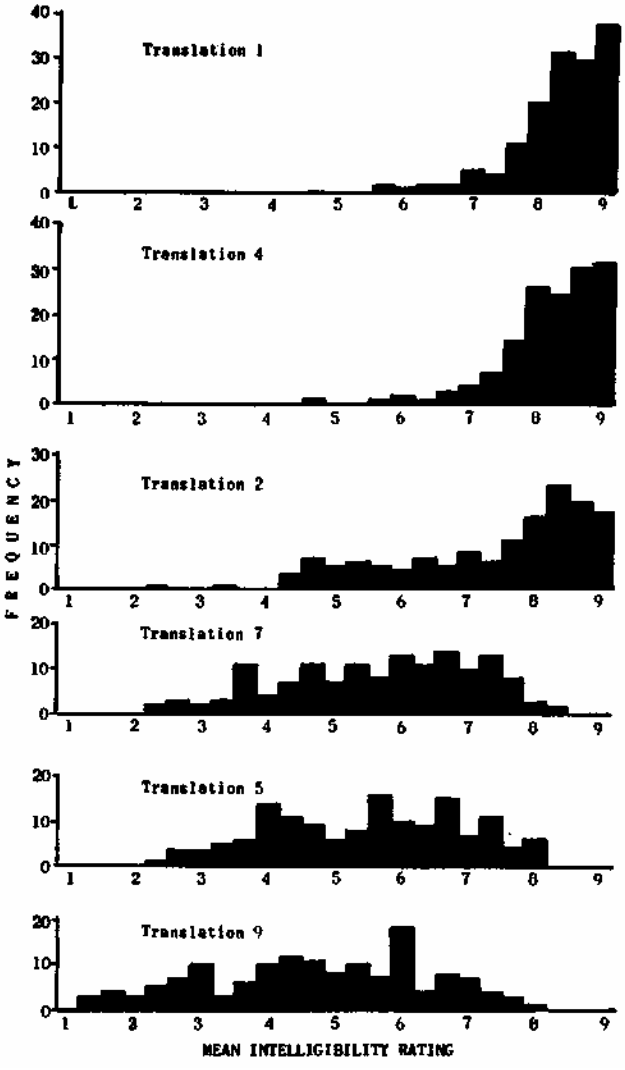

FIG. 1—Frequency distribution of monolinguals' mean intelligibility ratings of the 144 sentences in each of six translations. Translations 1, 4, and 2 are human translations; Translations 7, 5, and 9 are machine translations.

mean that the raters agree reliably that the sentences selected from a given passage in a given translation differ substantially, and further, that for any given passage, the translations are differentially effective for the different sentences. These findings agree with what we could have expected because it is obvious that machine-translation algorithms could be differentially successful for different kinds of sentences and lexical items. A detailed examination of the mean ratings for sentences (Fig. 1) shows, further, that sentences are much more variable in their intelligibility and informativeness when translated by machine than when translated by human translators. At least a few sentences translated by machine are indistinguishable from human translations, and it is tempting to add that at least a few sentences translated by humans look surprisingly like machine translations.

The within mean squares are estimates of the interrater variances, reflecting the degree to which the three raters of a given translation sentence differ in their ratings. For intelligibility and informativeness, they are (significantly) smaller in Part I of the experiment, using monolinguals; the converse is true, however, for time scores. The monolingual subjects, selected for high verbal intelligence and scientific interests, attained greater reliability in their ratings than did the Russian-reading subjects. In both parts of the experiment, the interrater variance is smaller for the intelligibility scale than it is for the informativeness scale; evidently the former is easier to make ratings on and produces more reliable ratings.

The over-all mean ratings and time scores shown in Table 4 give a concrete impression of the nature of the results. In terms of intelligibility, the three human translations are all fairly near the top of the scale, Translation No. 2 being the least acceptable of these. It is of interest to note that Translation No. 4, a "rapid" human translation, is nearly as high on the scale as Translation No. 1, the allegedly "careful," published, human translation. The three machine translations have average ratings near the middle of the scale and can as a whole be characterized by the phraseology attached to scale value 5 (see Table 1). Translation No. 9, an early attempt, is least intelligible.

The Russian readers tend to rate all translations a little higher in intelligibility, on the average, than do the monolingual raters; this is probably to be explained on the basis of the instructions to the Russian readers, which were to use any ingenuity or knowledge of Russian they might have to divine the meaning of the translations.

The rankings of the translations by the average ratings on the informativeness scale are almost precisely complementary to the rankings on intelligibility. Relative to the translations, the Russian readers tended to rate the originals at a slightly lower level of informativeness than the level at which the monolinguals rated the translated target sentences, but this is probably due to the fact that the Russian readers were better able to comprehend the translations by virtue of their knowledge of Russian word order and idiom. (The question of the translation adequacy of the target sentences rated by the monolinguals cannot be resolved from the present experiment. Because it was desired to preserve the symmetry of Parts I and II of the experiment, the Russian readers were not given the opportunity to evaluate the sentences of Translation No. 0 as translations of the Russian originals.)

The average reading-time scores show an almost perfect linear negative correlation with the average intelligibility ratings, and an almost perfect linear positive correlation with the informativeness ratings. The linearity of these relations strongly suggests that each of the two rating-scale variables used here can be regarded as being on an interval scale having equal units of measurement; they were established, of course, on the basis of the equal-appearing-intervals technique.

The Russian readers took slightly (but significantly) more time to comprehend the translation sentences than did the monolingual raters. Perhaps their knowledge of Russian allowed them or impelled them to study the translations more carefully, but perhaps, on the other hand, the results can be interpreted as showing that the monolinguals were quicker in comprehension by virtue of their greater scientific knowledge and interest.

It is worth pointing out that, for both the monolinguals and the Russian readers, the machine-translated sentences tended to take about twice as long to read and rate as the human-translated sentences.

The results displayed in Table 3 show only that, for each one of the three dependent variables in each part of the experiment, the means for the translations as shown in Table 4 differ so much that they could not reasonably have come from random sampling of the same population of observations. To test the significance of the differences between adjacent values when the means are ordered in magnitude, we use the Newman-Keuls test (Winer, 1962, pp. 80-85). The bracketings in Table 4 show the results of this test applied at the .01 level of significance to the ordered means. With respect to the mean values of every variable, all human translations are significantly different from all machine translations. Further, for most of the variables, human translation 2 is significantly inferior to human translations 1 and 4, and machine translation 9 is significantly inferior to machine translations 5 and 7. However, human translations 1 and 4 are in no case significantly different. Likewise, machine translations 5 and 7 are in no case significantly different in their mean values. It will be noted that the translations are generally better differentiated by ratings and performances of the monolinguals than by those of the bilinguals.

Discussion

The reader will doubtless have been struck by the high correlations among the three dependent variables used for evaluating translations in this study, even though, as noted above, they are conceptually independent. It must be pointed out, however, that high correlations are obtained only between *average* ratings for the translations, the averages being taken over raters, sentences, and passages. If the average ratings for sentences (always over three raters, in the present study) are examined, the correlations will not necessarily be extremely high. Numerous sentences can be found in the present data for which the locus of the average intelligibility and informativeness ratings on a two-dimensional plot falls considerably away from the locus of points for which intelligibility rating plus informativeness rating equals 10. It may be assumed that this phenomenon is not due solely to chance. Two

TABLE 6

TARGET SENTENCES, TRANSLATIONS, AND EVALUATIVE DATA FOR SENTENCE 8 IN PASSAGE 2, FOR PARTS I ("MONOLINGUAL") AND II ("BILINGUAL") OF THE TRANSLATION EXPERIMENT
(N = 3 Raters Each Sentence)

Target sentence (English version): What degree of automation now allows us to call a given mechanism an automaton?

Target sentence (original Russian): Какая степень автоматизации дает в настоящее время право назвать данный механизм автоматом?

TRANSLATION	PART	AVERAGE RATINGS: Intelligibility (A)	AVERAGE RATINGS: Informativeness (B)	A + B	AVERAGE TIME (secs.)
1. Careful (human): What degree of automation gives the right at present for a specific mechanism to be called an automaton?	I	8.00	1.67	9.67	7.00
	II	8.33	1.00	9.33	6.67
2. Quick (human): What degree of automation makes it right at the present time to call a given mechanism an automatic machine?	I	8.00	1.33	9.33	7.67
	II	8.67	1.00	9.67	5.33
4. Quick (human): What degree of automation presently bestows the right to call a certain piece of mechanism an automatic machine?	I	8.67	1.67	10.33	5.67
	II	8.67	1.33	10.00	8.00
5. Machine: What kind of degree of automation give/let at present right/law call given/data mechanism by automatic machine?	I	3.67	3.00	6.67	18.00
	II	6.33	3.33	9.33	9.33
7. Machine: Which degree of automation gives at present a right to call the given mechanism by an automatic device?	I	5.33	1.00	6.33	11.00
	II	8.00	1.67	9.67	12.00
9. Machine: Any/which/some/what degree/power of the automation gives into the present time/period the law/right to call the given mechanism by the automatic/slot mach. machine	I	5.00	7.67	12.67	24.67
	II	6.33	5.67	12.00	13.67

TABLE 7

TARGET SENTENCES, TRANSLATIONS, AND EVALUATIVE DATA FOR SENTENCE 10 IN PASSAGE 2, FOR PARTS I ("MONOLINGUAL") AND II ("BILINGUAL") OF THE TRANSLATION EXPERIMENT

(*N*= 3 Raters Each Sentence)

Target sentence (English version): However, by no means every machine may be called an automaton.

Target sentence (original Russian): Однако далеко не каждая машина называется автоматом.

TRANSLATION	PART	AVERAGE RATINGS: Intelligibility (A)	AVERAGE RATINGS: Informativeness (B)	A + B	AVERAGE TIME (secs.)
1. Careful (human): However, each machine is far from being called an automaton	I	8.33	5.67	14.00	4.33
	II	7.67	4.33	12.00	11.00
2. Quick (human): However, far from each machine is called an automatic machine	I	7.33	2.00	9.33	4.67
	II	4.00	4.33	8.33	21.67
4. Quick (human): However, it is not every machine that is referred to as an automatic machine	I	8.67	1.33	10.00	4.33
	II	9.00	1.33	10.33	5.33
5. Machine: However, by far not every machine is called automatic machine	I	7.00	2.00	9.00	7.00
	II	8.00	2.33	10.33	4.00
7. Machine: However far not each machine is called an automatic device	I	7.00	2.33	9.33	4.67
	II	7.67	3.00	10.67	10.67
9. Machine: However it far/far not each machine is called by the automatic/slot mach. machine	I	2.67	7.33	10.00	26.67
	II	6.00	2.33	8.33	9.00

TABLE 8

MAXIMUM LIKELIHOOD ESTIMATES OF TRUE VARIANCES (σ^2) FOR TRANSLATIONS, PASSAGES, SENTENCES, INTERACTIONS, AND ERROR FOR THREE DEPENDENT VARIABLES, BY TYPE OF RATER (M = MONOLINGUAL, B = BILINGUAL), DERIVED FROM THE PRESENT EXPERIMENT

SOURCE	MEAN RATINGS: Intelligibility M	MEAN RATINGS: Intelligibility B	MEAN RATINGS: Informativeness M	MEAN RATINGS: Informativeness B	MEAN READING TIMES PER SENTENCE: M	MEAN READING TIMES PER SENTENCE: B
Translation *(a)*	2.2747	2.0641	2.0150	2.1236	36.4706	30.9885
Passage *(b)*	[—.0082]*	.0045	[—.0273]*	.0104	[—.0678]*	.8110
Sentences (c)	.5141	.4145	1.0277	.5336	30.4790	35.5324
T X P *(ab)*	.0781	.0377	.0278	.0522	.0755	1.1673
T X S *(ac)*	.7928	.5053	1.6424	.9924	10.7230	23.8494
Error (e)	1.4133	1.7485	3.0753	3.2705	141.5769	93.4832

* These negative values may be replaced by zeros.

examples are to be found in Tables 6 and 7. For monolinguals, translations 5 and 7 in Table 6 are relatively unintelligible, and the target sentence is not very enlightening either. The converse case is illustrated by translation 1 in Table 8, where the translation seemed quite intelligible to both monolinguals and bilinguals. However, when they saw the target sentence (whether in English or in Russian), they perceived that it conveyed a rather different meaning from that conveyed by the translation. The translation can thus be regarded as somewhat inaccurate.

Over any sizable set of sentences in a given translation text, the tendencies for translations to be inaccurate or for the original sentences to be less than perfectly intelligible apparently counterbalance each other, with the result that there is an almost perfect negative correlation between average intelligibility rating and average informativeness rating. The correlation is slightly higher for monolinguals than for bilinguals. This would suggest that, for practical purposes, an entirely adequate method of evaluating human and mechanical scientific translations is simply to obtain intelligibility ratings of translation sentences from raters of high verbal intelligence and average them over raters and sentences. Our results indicate that if thirty-six sentences are selected at random from a translation and parceled out among eighteen raters in such a way that each sentence is rated by three raters (i.e., each rater rates six sentences), the sentences being interspersed among sentences from a varied collection of good, mediocre, and bad translations, then the standard error of the mean of all the ratings on the scale of intelligibility we have established will be about .17. This degree of precision should be sufficient to differentiate translations in most cases of practical importance. Surely it will serve to differentiate human from machine translations for a long time to come.

On the other hand, to guard against the possibility that a given translation source might be particularly subject to lack of fidelity, it would probably be desirable to obtain ratings not only on the intelligibility scale but also on the informativeness scale, and to note the extent to which the average ratings for the translation source tend to deviate from a position along the line defined by (intelligibility + informativeness) = 10.

The validity of the subjective ratings obtained in this experiment seems more or less self-evident, but it would be desirable to check it further by comparing the ratings with measurements obtained by other means, for example, by the reading-comprehension test method developed by Pfafflin (see reference 2).

A methodological detail that should be investigated further is the question of how important it is to screen raters for verbal intelligence and scientific knowledge.

Of greatest practical importance, however, would be a further investigation that would seek to establish the standard evaluation technique for which the present experiment sought to lay the foundations.

Suppose one had a set of sentences produced by a given translation source—a given human translator or a particular machine-translation program—and one wanted to obtain a mean rating for this translation source on one or both of the scales developed in the present experiment. Call these sentences the *probanda* sentences, or P-sentences. To employ the general procedure developed in the present experiment, it would be necessary to have available a set of translation sentences (with accompanying originals or translation equivalents) drawn from a variety of subject-matter sources and produced by a variety of translation systems, in such a manner that the mean ratings would fall approximately in flat (rectangular) distributions on the two rating scales. Call these the *comparanda* sentences, or C-sentences. It would then be necessary to set up a procedure whereby the P-sentences could be interspersed randomly among the C-sentences, the combined set to be rated by a panel of raters selected according to criteria to be specified, and trained suitably for the rating task. Most aspects of the process of arranging the materials to be rated and assembling and averaging the ratings could be programed for a computer and auxiliary equipment, such as optical scanners or machines for handling mark-sensing cards.

The questions that would remain to be answered in order to set up this procedure would be:

1. How many P-sentences from a given translation source should be rated in order to attain a given degree of sampling stability for the resultant mean ratings? How should these sentences be sampled from the output of the translation source?
2. How many C-sentences should be assembled in order to provide a minimally adequate "matrix" within which P-sentences could be interspersed? From how many different sources should these sentences be drawn?
3. How many raters would be required for the panel of raters in order to attain a given degree of precision for the resultant mean ratings for the translation source under study?

One can also conceive a situation in which it might be desirable to evaluate P-sentences from more than one translation source, in which case the answers to the above questions would become somewhat more complicated.

Some preliminary answers to these questions can be worked out from data collected in the present experiment. It is possible to solve the equations implied in Table 5 for estimates of true variance due to passages, sentences, and their interactions, and to set confidence bands for these estimates. These then can be used as a guide to estimating the degree of precision attainable through the use of a given number of P-sentences selected from a given number of disparate samples from a given translation source, rated by a given number of

raters. The estimates of true variance due to the various sources, for all three dependent variables used in the present experiment and for both monolingual and bilingual raters, are shown in Table 8.

Given n (the number of raters), q (the number of passages drawn from a translation source), and r (the number of sentences selected randomly from each passage), one can estimate the standard error of a mean value derived from nqr observations by the formula

$$\sigma_{\bar{Y}} = \sqrt{\frac{\sigma_e^2 + n\sigma_{as}^2 + nr\sigma_{ab}^2}{nqr}},$$

using the estimates of the pertinent variances in Table 8.

Received August 19, 1966

References

1. Miller, G. A., and Beebe-Center, J. G. "Some Psychological Methods for Evaluating the Quality of Translations," *Mechanical Translation,* Vol. 3 (1958), pp. 73-80.
2. Pfafflin, Sheila M. "Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments," *Mechanical Translation,* Vol. 8 (1965), pp. 2-8.
3. U.S. Dept. of Commerce, Office of Technical Services. *Machine and Thought: Excerpts.* Technical Translation TT 65-60307. Washington, D.C.: Government Printing Office, 1965.
4. Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill Book Co., 1962.